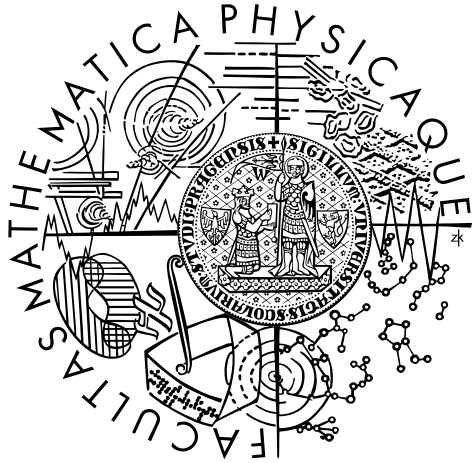
Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Tomáš Šurín

Zobrazovač Wikipedie pro mobilní zařízení

Katedra softwarového inženýrství

Vedoucí bakalářské práce: RNDr. Michal Žemlička, Ph.D.

Studijní program: Informatika, Programování

2010

Ďakujem RNDr. Michalovi Žemličkovi, Ph.D. za odborné vedenie mojej práce, za rady a za čas, ktorý mi behom vypracovania tejto práce venoval. Taktiež ďakujem ľuďom, ktorí mi pomohli s korektúrami.

Prehlasujem, že som svoju bakalársku prácu napísal samostatne a výhradne s použitím citovaných prameňov. Súhlasím s požičiavaním práce a jej zverejňovaním.

V Prahe dňa 26.07.2010 Tomáš Šurín

Obsah

Názov práce: Zobrazovač Wikipedie pro mobilní zařízení

Autor: Tomáš Šurín

Katedra (ústav): Katedra softwarového inženýrství

Vedúci bakalárskej práce: RNDr. Michal Žemlička, Ph.D.

e-mail vedúceho: Michal.Zemlicka@mff.cuni.cz

Abstrakt: Viacero ľudí považuje za veľmi praktické mať neustály prístup k encyklopedickým a iným referenčným informáciám. Toto je v súčasnosti veľmi dobre možné vďaka mobilným zariadeniam, ktoré umožňujú prístup k Internetu. Avšak existujú situácie, kedy nie je pripojenie k Internetu dostupné alebo nie je užívateľ za toto pripojenie ochotný platiť. Preto sme sa rozhodli umožniť užívateľom mobilných zariadení offline prístup k týmto informáciám. Konkrétne sa zaoberáme sprístupnením dát internetovej encyklopédie Wikipédia. Túto encyklopédiu sme zvolili ako zdroj dát pre náš projekt, pretože je veľmi obsiahla, má relatívne kvalitný obsah a ten je navyše voľne dostupný. Wikipédia beží na systéme MediaWiki, ktorý je používaný aj v množstve iných projektov. Navyše viacero firiem používa tento systém pre správu svojej vnútornej databáze znalostí. Pomocou popisovaného riešenia je preto možné sprístupniť aj dáta týchto projektov. Riešenie je navrhované s ohľadom na jednoduchú rozšíriteľnosť, pretože plánujeme pridať podporu aj iných informačných zdrojov.

Kľúčové slová: Wikipédia, Pocket PC, Windows Mobile, Offline

Title: Wikipedia viewer for mobile devices

Author: Tomáš Šurín

Department: Department of Software Engineering

Supervisor: RNDr. Michal Žemlička, Ph.D.

Supervisor’s e-mail address: Michal.Zemlicka@mff.cuni.cz

Abstract: Unlimited access to encyclopedic or other reference sources of information is considered by many people to be very convenient. Thanks to various mobile devices with Internet access, this is possible nowadays. There are, however, situations when there is no available Internet connection or the users are not willing to pay for it. Because of this we decided to allow offline access to this kind of information to users of mobile devices. Specifically we are trying to provide data from encyclopedia Wikipedia. We have chosen this encyclopedia because it is very comprehensive, it has good quality content and its data is freely available. Wikipedia runs on MediaWiki system, which is used in hundreds of other online projects. Additionally many corporations are using MediaWiki to manage their internal knowledgebase. Hence our solution can also provide data from these projects. Solution was designed with simple extensibility in mind as we are planning to add support for other information sources.

Keywords: Wikipedia, Pocket PC, Windows Mobile, Offline

1 Úvod

1.1 Motivácia

Žijeme v dobe, ktorej vládnu informácie. Valia sa na nás zo všetkých strán. Z klasických médií nám informácie už tradične poskytuje tlač, rádiové alebo televízne vysielanie. Na počiatku 90. rokov 20. storočia sa však do povedomia ľudí dostalo nové médium, ktoré sa stalo hlavným prostriedkom na získavanie a sprostredkúvanie informácií. Jedná sa o celosvetovú sieť Internet a hlavne o jej službu WWW.

Sprístupnenie cien a zrýchlenie pripojenia k Internetu na prelome tisícročí znamenalo masový záujem ľudí o túto technológiu. Pre mnohých sa Web stal hlavným a nenahraditeľným zdrojom informácií. Svedčí o tom aj penetrácia internetových prípojok v domácnostiach, školách, firmách či úradoch.

Jedným z najobľúbenejších a najznámejších webových zdrojov informácií zo všetkých oblastí je encyklopédia Wikipédia. Jej obsah je vytváraný samotnými užívateľmi a je ponúkaný zadarmo všetkým záujemcom. Wikipédia predstavuje kvalitný zdroj informácií – hlavne čo sa týka jej rozsahu a miery pokrytia tém.

Problémom však je, že Wikipédia je dostupná len ak máme k dispozícii pripojenie k internetu. Čo však robiť, ak vznikne potreba prístupu k Wikipédii bez pripojenia? Napríklad ak je pripojenie nedostupné alebo je cenovo nevýhodné? Riešením môže byť systém, ktorý by poskytoval možnosť vytvoriť lokálne úložisko dát z encyklopédie a pristupovať k nim offline.

Použitie tohto systému nie je obmedzené len na stolné počítače alebo notebooky. V posledných rokoch sa začali rozširovať mobilné zariadenia s vlastným operačným systémom – tzv. smartphony alebo PDA. Tieto ponúkajú potrebný výkon a vďaka relatívne nízkym cenám pamäťových kariet, taktiež potrebné úložisko pre nasadenie vyššie uvedeného systému. Takýto mobilný systém s dátami z anglickej encyklopédie wikipédia sa približuje k fiktívnemu „Stopárovmu sprievodcovi galaxiou“[1], či už to je rozsahom obsiahnutých informácií, alebo ich dôveryhodnosťou.

Takáto mobilná možnosť prístupu k encyklopedickým informáciám alebo všeobecne k akýmkoľvek informačným zdrojom má veľké čaro. Už nie sme obmedzení len na

veľké zariadenia s nízkou prípadne žiadnou mobilitou alebo knižnými zdrojmi s vysokou hmotnosťou. Väčšinu potrebných informácií môžeme mať stále so sebou vo vlastnom vrecku.

1.2 Ciele

Cieľom práce je implementovať program, ktorý bude sprístupňovať textový obsah internetovej encyklopédie Wikipédia k offline prehliadaniu na zariadeniach s operačným systémom Windows Mobile.

Cieľom je taktiež navrhnúť program tak, aby bol jednoducho rozšíriteľný.

Ďalším cieľom je navrhnúť vhodný spôsob kompresie dátových súborov.

1.3 Štruktúra práce

Riešenie, ktoré je predmetom práce, pozostáva z 3 samostatných častí:

1. **Zobrazovacia časť** - Jej úlohou je sprostredkovať články z Wikipédie užívateľovi na mobilnom zariadení. Túto časť tvorí program WikiReader.
2. **Konverzná časť** – Jej úlohou je vytvorenie dátových súborov pre zobrazovaciu časť. Túto časť tvorí program WikiConvert.
3. **Kompresná časť** - Jej úlohou je kompresia dátových súborov. Túto časť tvorí program WZip.

Samotná práca sa skladá z týchto častí:

- požiadavky na riešenie (2. kapitola)
- popis existujúcich implementácií (3. kapitola)
- všeobecné informácie o Wikipédii (4. kapitola)
- popis a analýza databázových exportov zo systému MediaWiki (5. kapitola)
- analýza kompresných metód a výber najvhodnejšej metódy (6. kapitola)

- popis dátových súborov, postup ich vytvorenia z databázových exportov pomocou programu WikiConvert a popis jeho implementácie (7. kapitola)
- užívateľská dokumentácia k programu WikiReader (8. kapitola)
- popis implementácie programu WikiReader (9., 10. a 11. kapitola)
- popis možných rozšírení riešenia (12. kapitola)

Diagram 1 znázorňuje princíp práce riešenia popisovaného v práci.

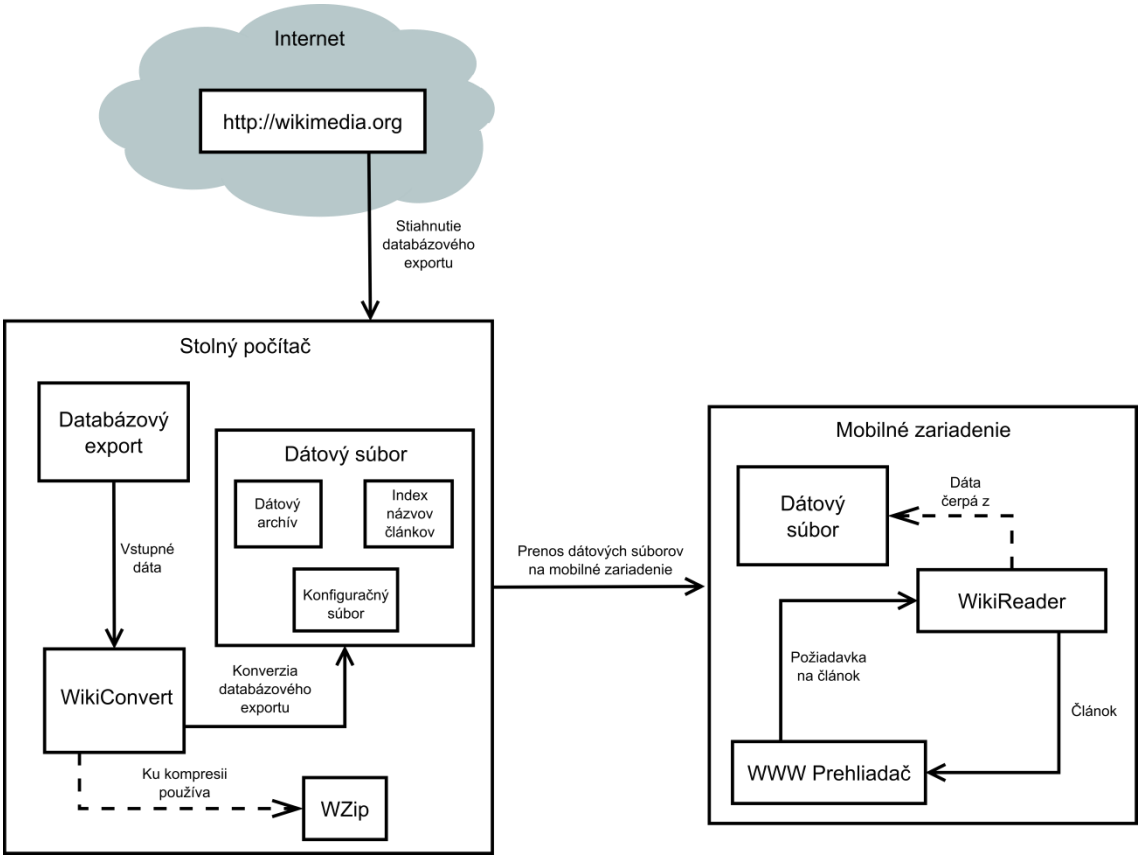


Diagram 1: Hrubá schéma riešenia popisovaného v tejto práci

2 Požiadavky

V tejto kapitole sú zhrnuté požiadavky na riešenie tak, ako boli zhromaždené na začiatku projektu a podľa ktorých boli vytvárané príslušné aplikácie.

2.1 Požiadavky na všetky časti riešenia

Chceme aby bolo riešenie voľne šíriteľné a toto musia dovoľovať aj použité časti. Preto od nich chceme, aby boli voľne dostupné, použiteľné a šíriteľné. Tieto časti by mali mať zaistenú alebo aspoň umožnenú údržbu (mali by za nimi stáť programátorské tímy, alebo by aspoň mali byť open-source).

2.2 Zobrazovacia časť

Zobrazovacia časť bude určená pre mobilné zariadenia. Je dôležitá jej nízka pamäťová a výkonová náročnosť, pretože mobilné zariadenia disponujú malým množstvom pamäte a zbytočné plytvanie výkonom sa podpisuje na menšej výdrži batérie. Pre nedostatok času vystačíme so zariadeniami s operačným systémom Windows Mobile.

Zobrazovacia časť by mala byť schopná okrem obsahu encyklopédie Wikipédia poskytovať aj obsah iných projektov používajúci systém WikiMedia ako systém pre správu obsahu (content management system).

Zobrazovacia časť by mala poskytovať čo najlepšiu podporu formátovania článkov. Zobrazovanie článkov by malo byť navyše dostatočne rýchle – tj. aby doba zobrazenia článku bola pre užívateľa únosná.

Riešenie plánujeme v budúcnosti rozšíriť o nové možnosti, napríklad o podporu iných informačných zdrojov. Preto by mala zobrazovacia časť podporovať mechanizmy umožňujúce jednoduchú rozšíriteľnosť – napríklad podporu zásuvných modulov.

2.3 Konverzná časť

Konverzná časť bude určená pre stolný počítač. Mala by byť čo najmenej závislá na použitom operačnom systéme.

Na rýchlosti konverzie nám veľmi nezáleží, avšak mala by byť znesiteľná.

Informácií je veľké množstvo a kladú vysoké nároky na dátové úložisko. Preto je vhodné, aby dátové súbory dosahovali čo najmenšiu veľkosť.

2.4 Kompresná časť

Kompresný formát zvolený pre kompresiu dátových súborov by mal ponúkať vysoký kompresný pomer a dostatočne rýchlu dekompresiu na mobilnom zariadení.

Taktiež je vhodné, aby bol kompresný formát dostatočne odolný proti prenosovým chybám a aby umožňoval overenie integrity dát (kontrolné súčty).

3 Existujúce implementácie

V tejto kapitole sú popísané najznámejšie softwarové riešenia umožňujúce offline prehliadanie Wikipédie. Dôraz je kladený na riešenia pre systém Windows Mobile.

3.1 TomeRaider

TomeRaider [2] je program pre viacero mobilných platforiem vrátane Windows Mobile a pre Windows. Program slúži ako čítačka elektronických kníh. Špeciálne je však vhodný pre čítanie veľkých databází informácií, akými sú napríklad Wikipédia. Program funguje na princípe zobrazovania statického HTML popísaného v kapitole 5.4.2.

Vyskúšali sme dátový súbor obsahujúci textový obsah z anglickej Wikipédie, ktorý bolo možné získať tu [3][1].

Výhodou programu TomeRaider je dobrá kompresia a výborná podpora formátovania. Dátový súbor s anglickou Wikipédiou, ktorý sme testovali, podporuje tabuľky, zoznamy, šablóny, matematické vzorce, kódovanie unicode atď. Niektoré verzie dátových súborov dokonca podporovali obrázky.

Nevýhodou programu TomeRaider je hlavne obmedzenie veľkosti súborov na 4GB, čo znemožňuje vytvorenie dátových súborov súčasnej verzie anglickej Wikipédie. Ďalšou nevýhodou je, že je platený.

3.2 Mdict

Mdict [4] je freeware slovníkový program pre Windows Mobile a Windows. Okrem dátových súborov obsahujúcich rôzne slovníky preň existujú aj dátové súbory obsahujúce dáta z viacerých jazykových mutácií Wikipédie (je ich možné stiahnuť napríklad na [5]).

Jeho hlavnou nevýhodou je v prípade zobrazovania offline obsahu Wikipédie, minimálna podpora formátovania (žiadne tabuľky alebo zoznamy ...).

[1] V súčasnosti už nie je možné z tohto zdroja získať dátové súbory Wikipédie pre program TomeRaider.

3.3 WikiPock

WikiPock [6] je program na offline prístup k dátam z Wikipédie pre mobilné zariadenia s operačným systémom Windows Mobile, BlackBerry a Android.

Nevýhodou programu je, že je platený a navyše neexistuje žiadna demo verzia na otestovanie funkčnosti.

3.4 Riešenia pre iné operačné systémy

3.4.1 Okawix

Okawix [7] je open source projekt pre Windows, Mac OS a Linux. Poskytuje offline prístup k veľkému množstvu projektov organizácie Wikimedia Foundation.

Výhodou programu je výborná podpora formátovania veľmi podobná originálnej online Wikipédii. Navyše je pod licenciou GPL.

Nevýhodou programu je obmedzenosť na vyššie uvedené platformy.

3.4.2 Wiki2Touch

Wiki2Touch [8] je offline čítačka Wikipédie pre iPhone. Princíp fungovania tohto programu je veľmi podobný programu WikiReader - články ponúka prostredníctvom HTTP serveru a pracuje na princípe dynamického generovania HTML popísaného v kapitole 5.4.3.

Nevýhodou programu je, že je len pre iPhone. Výhodou je výborná podpora formátovania so všetkým čo k tomu patrí – tabuľky, matematické vzorce, šablóny atď.

4 Všeobecné informácie o projekte Wikipédia

V tejto kapitole sú uvedené základné informácie o organizácii Wikimedia Foundation, Inc. a jej projekte internetovej encyklopédie Wikipédia.

4.1 Základné informácie

Wikipédia je voľná internetová encyklopédia umožňujúca komukoľvek vytvárať a upravovať jej obsah.

Wikipédia je projekt zastrešovaný neziskovou organizáciou Wikimedia Foundation, Inc. Cieľom organizácie Wikimedia Foundation je starať sa o vývoj a propagáciu nástrojov umožňujúcich vytváranie wiki projektov a poskytovanie ich obsahu verejnosti zdarma [9]. Medzi jej najznámejšie wiki projekty patrí už vyššie spomínaná encyklopédia Wikipédia spolu s jej jazykovými mutáciami, slovník Wiktionary alebo databáza citácií Wikiquote.

Wiki projekt/systém predstavuje webový portál, ktorý svojím užívateľom poskytuje možnosť jednoduchej editácie svojho obsahu. Obsah celého systému je delený na články, ktoré odpovedajú jednotlivým čiastočným stránkam portálu. Väčšinou poskytuje taktiež správu verzií jednotlivých článkov.

4.2 Základné princípy

Wikipédia je wiki systém, ktorý má nasledujúce vlastnosti [10]:

- Každý užívateľ môže editovať jej obsah. Výnimku tvoria iba články chránené proti vandalizmu – jedná sa hlavne o články popisujúce citlivé témy, napríklad články týkajúce sa viery, sexuálnej orientácie alebo politiky.
- Editovaný obsah je ihneď dostupný verejnosti. Jeho vlastníkom je komunita editorov.
- Pri editácii používa vlastný značkovací jazyk – tzv. wiki markup. Má to výhodu v tom, že prispievajúci užívateľ nemusí poznať jazyk HTML. Navyše je možné používať aj HTML alebo CSS, pričom však platí, že potenciálne nebezpečné

elementy jazyka HTML, akým je napríklad tag `<script>`, systém pri zobrazovaní stránky odstráni.

4.3 Systém MediaWiki

MediaWiki je systém pôvodne vytvorený pre potreby Wikipédie. V súčasnosti je používaný aj v ostatných projektoch Wikimedia Foundation a v množstve iných webových projektov, ako systém pre správu obsahu. MediaWiki je licencovaný pod GPL [11]. Je napísaný v PHP a pre ukladanie dát používa databázu MySQL [12].

Vzhľadom na rovnaký princíp fungovania a rovnaký formát dát všetkých projektov používajúcich systém MediaWiki je možné použiť riešenie popisované v tejto práci nielen pre offline prístup k Wikipédii, ale taktiež k iným projektom používajúcim systém MediaWiki. Riešenie je napríklad možné použiť pre offline prístup k dátam z wiki systémov, ktoré sú používané vo viacerých firmách pre správu dokumentácie. Jedinou podmienkou použitia je dostupnosť dát[1] z daného projektu. V ďalšom texte sa však budeme venovať len encyklopédii Wikipédia.

4.4 Výhody a nevýhody

Veľkou výhodou encyklopédie Wikipedia je rýchlosť jej rastu a vysoká miera pokrytia tém. Navyše obsahuje aj viacero neobvyklých tém, ktoré by sa do klasických encyklopédií nedostali.

Hlavné nevýhody Wikipédie plynú z kolaboratívneho prístupu k vytváraniu jej obsahu. Jedná sa napríklad o viaceré problémy s dôveryhodnosťou, vandalizmom a úmyselným vložením nesprávnej informácie.

[1] Tj. databázových exportov

5 Získanie obsahu

V tejto kapitole je popísaný spôsob získania obsahu encyklopédie Wikipédia a analýza týchto vstupných súborov.

5.1 Stiahnutie databázových exportov

Wikipédia ponúka svoj textový obsah voľne k stiahnutiu vo forme databázových exportov systému MediaWiki. Tieto exporty obsahujú jednotlivé články projektu vo wiki syntaxi. V ďalšom texte bude pod pojmom databázový export myslený export dát zo systému MediaWiki. Obsiahnutý textový obsah je pod licenciou Creative Commons Attribution-ShareAlike 3.0 [13] a GFDL [14].

Obrázky a iný multimediálny obsah sú však pod inými licenciami. V súčasnosti Wikipédia neponúka žiadnu možnosť ako stiahnuť všetky obrázky prípadne iné multimediálne súbory. Jedinou možnosťou, ako ich získať, je ich manuálne stiahnutie napríklad pomocou skriptu. Táto možnosť však kladie vysoké nároky na servery Wikipédie a na dátové úložisko. Preto riešenie používané v tejto práci nepodporuje zobrazovanie obrázkov a iných multimediálnych súborov. Namiesto toho poskytuje online odkaz na daný súbor.

Databázové exporty projektov WikiMedia Foundation je možné stiahnuť z [15]. Špeciálne pre anglickú wikipédiu sa jedná o projekt s názvom „*enwiki*“, pre českú „*cswiki*“, pre slovenskú „*skwiki*“ atď. Pre potreby offline čítačky Wikipédie je potrebné stiahnuť súbor s názvom končiacim „*pages-articles.xml.bz2*“, ktorý obsahuje len najnovšiu revíziu článkov a neobsahuje užívateľské stránky a diskusiu.

Ďalšou možnosťou získania dát zo systému MediaWiki je použitím stránky wiki projektu s názvom „*Special:Export*“. Táto stránka umožňuje vytvorenie databázového exportu pozostávajúceho len z vybraných článkov alebo kategórií.

5.2 Formát databázových exportov

Databázové exporty systému MediaWiki sú vo formáte XML v kódovaní UTF-8.

Pozostávajú z 2 základných častí:

- Informácie o projekte, z ktorého bol export vytvorený. Táto časť taktiež obsahuje lokalizované prefixy menných priestorov.
- Stránky predstavujúce jednotlivé články vo wiki syntaxi. Každá stránka obsahuje taktiež viaceré metainformácie, ako sú napríklad informácie o revízii.

Cesta k XML schéme databázového exportu je uvedená na jeho začiatku. Je ju možné nájsť v [16].

5.3 Menné priestory

Články v systéme MediaWiki sú rozdelené do tzv. menných priestorov. Tieto umožňujú separáciu rôznych druhov článkov v rámci systému. Jedná sa napríklad o encyklopedické články, články s informáciami o obrázkoch alebo články tvoriace diskusiu. Úplný názov článku je potom tvorený prefixom menného priestoru, dvojbodkou a samotným názvom článku.

Prefixy menných priestorov nie sú jednoznačné v rámci rôznych projektov. Sú špecifikované na začiatku databázového exportu v časti informácie o projekte. Napríklad menný priestor reprezentujúci šablóny má v anglickej Wikipédii prefix „*Template*“, v slovenskej zase „*Šablóna*“. Úplný názov článku „*Demo*“ nachádzajúcom sa v tomto mennom priestore je potom „*Template:Demo*“ respektíve „*Šablóna:Demo*“.

Jednoznačne sú menné priestory identifikovateľné medzi rôznymi projektmi ich číselným identifikátorom. Napríklad hlavný menný priestor má identifikátor 0, šablóny 10, kategórie 14.

Menným priestorom, v ktorom sa nachádzajú encyklopedické články, je hlavný menný priestor. Tento priestor nemá žiadny prefix.

5.4 Možnosti offline prehliadania obsahu Wikipédie

5.4.1 Lokálny server

Jednou z možností offline prehliadania Wikipédie je použitie inštancie MediaWiki bežiacej na lokálnom serveri s importovaným databázovým exportom. Systém MediaWiki je možné stiahnuť z [17]. Postup importovania dát je popísaný v [18].

Problémom tohto riešenia je podľa [18] v tom, že importovanie dát môže trvať dlhý čas - v prípade anglickej Wikipédie aj viac dní. V prípade systému Windows Mobile je ďalším problémom, že sme neboli schopný nájsť riešenie pre beh HTTP serveru s podporou PHP a MySQL.

5.4.2 Statické HTML

Ďalšou z možností offline prehliadania je vytvorenie dátových súborov, obsahujúcich články vo forme HTML súborov, ktoré boli vygenerované z databázových exportov.

Nevýhodou tohto riešenia je, že HTML stránky majú väčšiu veľkosť než dáta vo wiki jazyku. Navyše konverzia dát taktiež trvá nezanedbateľný čas. Výhodou je, že netreba nič za behu generovať ako v prípade dynamického generovania HTML.

5.4.3 Dynamické generovanie HTML

Ďalšou možnosťou offline prehliadania je dynamické generovanie článku vo forme HTML súborov z dát vo wiki jazyku za behu programu.

Toto riešenie dosahuje najmenšiu veľkosť dátových súborov zo všetkých spomínaných riešení. Vytvorenie dátových súborov taktiež trvá najkratšie. Ďalšou výhodou je, že v prípade zmeny netýkajúcej sa formátu dátových súborov netreba distribuovať užívateľom nové dátové súbory ale iba aktualizáciu programu. Nevýhodou však je, že treba za behu generovať HTML, čo sa podpisuje na pomalšom zobrazovaní stránok a vyššom zaťažení procesoru v mobilnom zariadení.

Kvôli požiadavke čo najmenšej veľkosti dátových súborov a subjektívnych dôvodov je toto riešenie použité aj v tejto práci. Navyše predpokladáme aspoň z počiatku malú komunitu užívateľov a preto je toto riešenie výhodné – veľa článkov nebude nikdy prehliadaných. Časť, ktorá má za úlohu konverziu z wiki jazyka do HTML navyše nie je vo finálnej verzii a pri použití statického HTML by bolo pri každej zmene tejto časti potrebné vytvoriť nové dátové súbory a distribuovať ich užívateľom.

5.5 Analýza vstupných dát

Pred začatím práce na programoch WikiConvert a WikiReader bolo potrebné vykonať analýzu obsahu databázových exportov. Cieľom bolo zistiť podiel veľkostí jednotlivých

menných priestorov a podiel článkov jednotlivých dĺžok. Analýza bola vykonaná na databázovom exporte anglickej Wikipédie „*enwiki-20100130-pages-articles.xml*“.

5.5.1 Metodika merania

Na analýzu bol použitý program WikiStatistics, nachádzajúci sa na priloženom DVD. Jedná sa o jednoduchý parser databázových exportov napísaný v jazyku Java, ktorého zdrojový kód tvoril základ pre program WikiConvert.

5.5.2 Podiel menných priestorov

Diagram 2 znázorňuje podiel veľkostí jednotlivých menných priestorov v databázovom exporte.

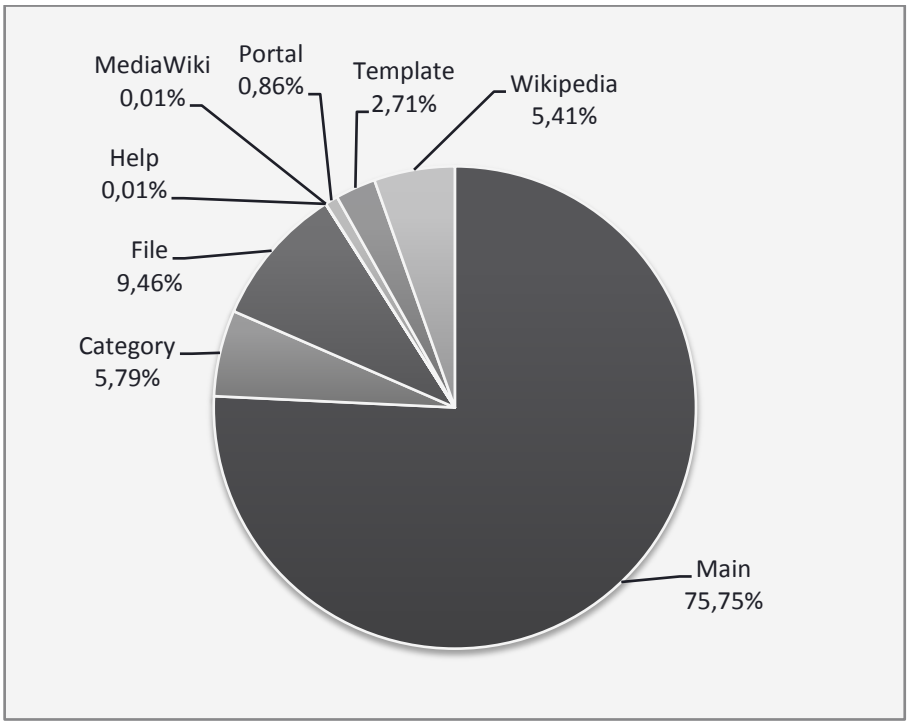


Diagram 2: Podiel menných priestorov anglickej Wikipédie

Najdôležitejším menným priestorom je hlavný menný priestor - „*Main“, ktorý* obsahuje encyklopedické články. Z analýzy vyplýva, že ostatné menné priestory prispievajú nezanedbateľným spôsobom k veľkosti databázového exportu. Pre potreby zmenšenia výsledného dátového súboru bude preto môcť užívateľ pri konverzii špecifikovať, ktoré menné priestory si želá vynechať (viď kapitola 6.2).

5.5.1 Dĺžky jednotlivých článkov

Diagram 3 znázorňuje rozdelenie dĺžok článkov v bajtoch. Jednotlivé dieliky na x-ovej osi reprezentujú dĺžku článkov väčšiu ako číslo dieliku a menšiu alebo rovnú ako číslo nasledujúceho dieliku. Y-ová os predstavuje počet článkov.

Najdlhší článok má dĺžku 405 338 bajtov.

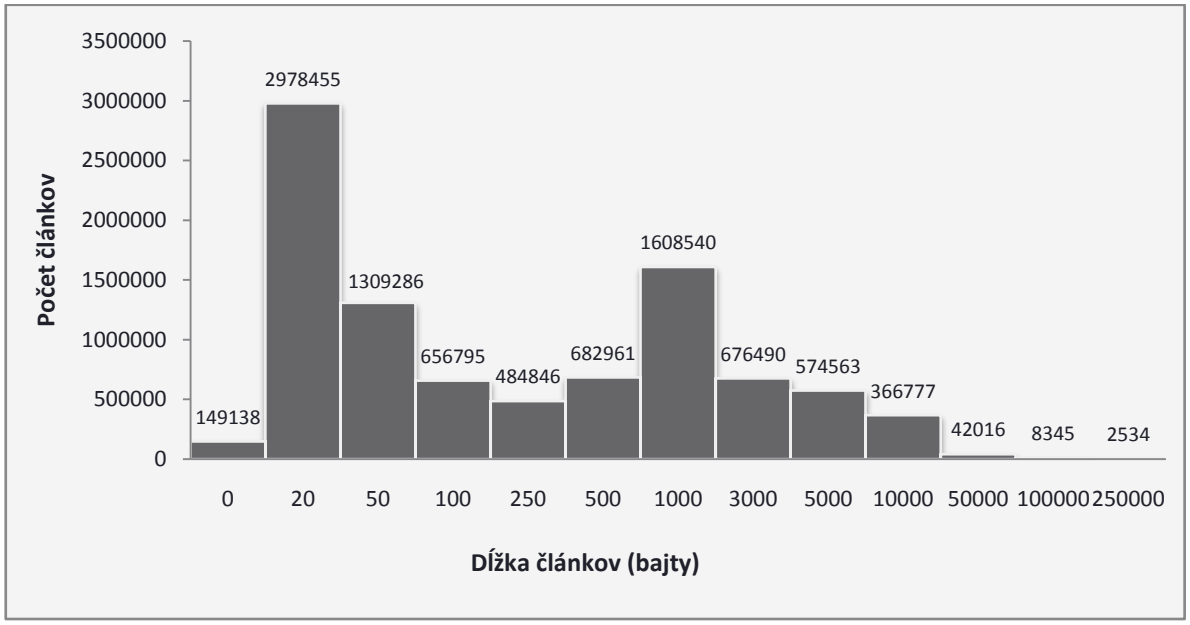


Diagram 3: Dĺžka článkov anglickej Wikipédie

Analýza dĺžok článkov bola potrebná pre zistenie vhodnej veľkosti bloku pri kompresii dátových súborov.

6 Kompresia

V tejto kapitole je popísaný spôsob výberu vhodnej kompresnej metódy používanej v dátových súboroch.

6.1 Uvažované kompresné metódy

Pri výbere kompresnej metódy bol kladený dôraz na to, aby existovala open source implementácia daného formátu a aby ponúkal použitý algoritmus rýchlu dekompresiu. Ďalším dôležitým faktorom bola aj veľkosť výsledného komprimovaného súboru. Samozrejme sa jedná o bezstratové kompresné formáty.

K testovaniu boli vybrané nasledujúce formáty:

- 7z [19] - Podporuje viacero kompresných algoritmov. Testovaný bol jeho základný algoritmus LZMA. LZMA poskytuje vysoký kompresný pomer. Rýchlosť kompresie a dekompresie je taktiež relatívne dobrá.

- zip [20] - Používa kompresný algoritmus Deflate. Tento algoritmus je používaný v širokom spektre aplikácií. Výhodou je, že existuje veľké množstvo implementácií algoritmu Deflate a formátu zip.

- gzip [21] – Jedná sa o implementáciu algoritmu Deflate. Umožňuje komprimovať len 1 súbor[1]. Výhodou formátu gzip je, že je súčasťou špecifikácie HTTP/1.1. Preto je možné klientovi, ktorý podporuje HTTP/1.1, poslať HTTP odpoveď skomprimovanú touto metódou. Túto vlastnosť by bolo možné využiť pri staticky generovanom HTML (viď kapitola 5.4.2), kde by sa o dekompresiu staral až klient.

- bzip2 [22] – Umožňuje komprimovať len 1 súbor[1]. Používa algoritmus bzip2. Databázové exporty Wikipédie sú ponúkané na stiahnutie práve v tomto formáte. Ponúka sa preto možnosť vytvorenia dátových súborov tak, že sa vytvorí iba index názvov článkov.

[1] Pre kompresiu viacerých súborov sa v praxi používa v kombinácii s utilitou tar.

K testovaniu sme chceli zaradiť aj kompresné formáty používajúce algoritmy slovnej a slabikovej kompresie [23], avšak nebolo možné dokončiť testy pomocou dostupných implementácií týchto metód.

6.2 Testy

Pre zistenie vhodnosti konkrétneho kompresného formátu pre použitie v dátových súboroch boli vykonané 2 jednoduché testy.

6.2.1 Metodika testovania

Prvý test bol zameraný na zistenie hrubých výsledkov. Databázový export „*simplewiki-20090912-pages-articles.xml*“ bol skomprimovaný na stolnom počítači a následne dekomprimovaný na mobilnom zariadení so systémom Windows Mobile. Sledované boli čas kompresie, kompresný pomer, čas dekompresie a približné využitie pamäte pri dekompresii.

Kompresia bola vykonaná na počítači s procesorom Core2Duo na frekvencii 2 GHz. Pri kompresii boli použité implicitné nastavenia kompresného programu. Kompresné programy použité na kompresiu jednotlivých formátov sú uvedené v tabuľke 1. Dekompresia bola vykonaná na zariadení HP iPAQ 214 s použitím programu 7-Zip 9.07 beta.

V druhom teste boli testované už len formáty 7z a zip[1]. Program WikiConvert bol použitý na vyššie spomenutom databázovom exporte. Pri konverzii bol nastavený limit veľkosti blokov na 50 KB, bola pri nej vypnutá solid archivácia a boli použité implicitné nastavenia kompresného programu.

6.2.2 Výsledky testov

Namerané výsledky 1. testu sú uvedené v tabuľke 1. Pre porovnanie bol test vykonaný aj na jednom z najznámejších zástupcov komerčných kompresných formátov/algoritmov – formáte rar.

Formát / algoritmus	Čas kompresie (stolný počítač)[2]	Kompresný pomer[3]	Čas dekompresie (mobilné zariadenie)[2]	Využitie pamäte pri dekompresii	Kompresný program
7z/LZMA	1:57	19,18%	1:28	9.5 MB	7z 4.65

[1] Bzip2 a gzip nepodporuje komprimáciu viacerých súborov.
[2] Čas je vo formáte minúty:sekundy.
[3] Pomer veľkosti výsledného archívu k veľkosti pôvodného databázového exportu.

zip/Deflate	0:16	28,49%	2:30	200 KB	Info-Zip 2.3
gzip/Deflate	0:46	28,50%	2:19	200 KB	gzip 1.2.4
bzip2/Bzip2	0:45	21,42%	4:45	3.5 MB	bzip2 1.0.5
rar/Rar	0:55	21,57%	2:05	5 MB	rar 3.90

Tabuľka 1: Výsledky testu kompresných metód

Najlepší kompresný pomer spomedzi testovaných kompresných metód dosahuje formát 7z. Paradoxne ponúka aj najrýchlejšiu dekompresiu na mobilnom zariadení. Jeho nevýhodou je pomalá kompresia a veľká spotreba pamäti pri dekompresii. Táto spotreba pamäti sa však odvíja od veľkosti používaného slovníka[1]. V teste bol používaný slovník o veľkosti 8 MB.

Algoritmus LZMA dovoľuje používať slovníky začínajúce už na veľkosti 4 KB. Takže spotreba pamäte je veľmi dobre škálovateľná. V tabuľke 2 sú výsledky 1. testu pre rôzne veľkosti používaného slovníka. Z testu je vidieť, že algoritmus LZMA dosahuje konkurenčne schopné výsledky aj pri použití menšieho slovníka.

Veľkosť slovníka (KB)	Čas kompresie (stolný počítač)[2]	Kompresný pomer[3]	Čas dekompresie (mobilné zariadenie)[2]
4096	0:56	28,91%	4:12
8192	0:54	27,04%	3:06
16 384	0:54	25,77%	2:45
32 768	0:55	24,81%	2:35
65 536	0:55	23,96%	1:57
131 072	0:57	23,16%	1:52
262 144	0:59	22,36%	1:40
524 288	1:00	21,61%	1:49
1 048 576	1:08	20,92%	1:40
2 097 152	1:21	20,27%	1:43
4 194 304	1:38	19,70%	1:32
8 388 608	1:52	19,18%	1:28

Tabuľka 2: Výsledky testu 7z v závislosti na veľkosti slovníka

Namerané výsledky 2. testu sú uvedené v tabuľke 3.

Formát	Čas konverzie (stolný počítač)[2]	Kompresný pomer[3]	Čas dekompresie (mobilné zariadenie)[2]	Kompresný program
7z	2:24	21,89%	11:00	7z 4.65
zip	2:05	24,85%	11:08	Info-Zip 2.3

Tabuľka 3: Výsledky testu kompresných metód v programe WikiConvert

[1] Podľa dokumentácie priloženej k LZMA SDK [24] potrebuje LZMA pre dekompresiu iba pamäť pre slovník a 16 KB pre vnútornú stavovú štruktúru.
[2] Čas je vo formáte minúty:sekundy.
[3] Pomer veľkosti výsledného dátového archívu k veľkosti pôvodného databázového exportu.

6.3 Zvolený formát

S prihliadnutím k výsledkom vyššie uvedených testov bol zvolený formát 7z. Avšak tento formát sa neskôr ukázal ako neefektívny. Problémom bola nemožnosť náhodného prístupu k ľubovoľnému komprimovanému súboru. Navyše implementácia dekompresie programom 7z si informácie o jednotlivých súboroch uchováva v pamäti, čo znamená nezanedbateľnú pamäťovú náročnosť. Preto bol vytvorený nový kompresný formát, ktorý používa algoritmus LZMA. Popis tohto formátu sa nachádza v kapitole 7.1.2. Pre kompresiu a dekompresiu tohto formátu bol vytvorený program WZip.

6.4 WZip - implementácia

Zdrojový kód programu WZip je napísaný v jazyku C. Jedná sa o modifikáciu programu LzmaUtil, ktorý sa nachádza v LZMA SDK [24]. Program LzmaUtil umožňuje kompresiu a dekompresiu súboru algoritmom LZMA. Výsledkom je LZMA blok vo formáte popísanom v kapitole 7.1.2.

Program WZip umožňuje komprimovať viacero súborov zároveň. Každý súbor ukladá do samostatného LZMA bloku. Na rozdiel od formátu 7z neobsahuje archív formátu WZip názvy súborov a iné metainformácie. Namiesto toho obsahuje pole ofsetov jednotlivých LZMA blokov.

7 Vytvorenie dátových súborov

V tejto kapitole je popísaná implementácia programu WikiConvert, spôsob jeho použitia na vytvorenie dátových súborov a formát týchto súborov.

7.1 Formát dátových súborov

Program WikiReader potrebuje pre zobrazovanie obsahu Wikipédie dáta vo vhodnom formáte. Tieto dáta je možné vytvoriť na klasickom stolnom počítači pomocou programu WikiConvert z databázového exportu wiki projektu. Výsledkom konverzie dát wiki projektu programom WikiConvert sú nasledujúce súbory:

- **konfiguračný súbor** – nastavenia prislúchajúce k danému wiki projektu.
- **dátový archív** – vhodným spôsobom skomprimované dáta článkov vo wiki značkovacom jazyku. Môže ho tvoriť viacero fyzických súborov operačného systému (viď koniec kapitoly 7.1.2).
- **index názvov článkov** – dátová štruktúra umožňujúca rýchle vyhľadávanie adresy článku v skomprimovaných dátach podľa jeho názvu. Taktiež umožňuje vyhľadávanie na čiastočnú zhodu s názvom článku[1].

Tieto súbory predstavujú jeden dátový súbor. V programe WikiReader je možné používať takýchto dátových súborov viacero súčasne. Napríklad je možné mať zároveň otvorené dátové súbory anglickej, českej a slovenskej Wikipédie.

7.1.1 Konfiguračný súbor

Konfiguračný súbor obsahuje nastavenia projektu. Jeho formát je popísaný v prílohe C.

7.1.2 Dátový archív

Formát dátového archívu je znázornený na diagrame 4.

[1] V programe je implementované vyhľadávanie na čiastočnú zhodu od začiatku názvu (tj. nie vo vnútri názvu).

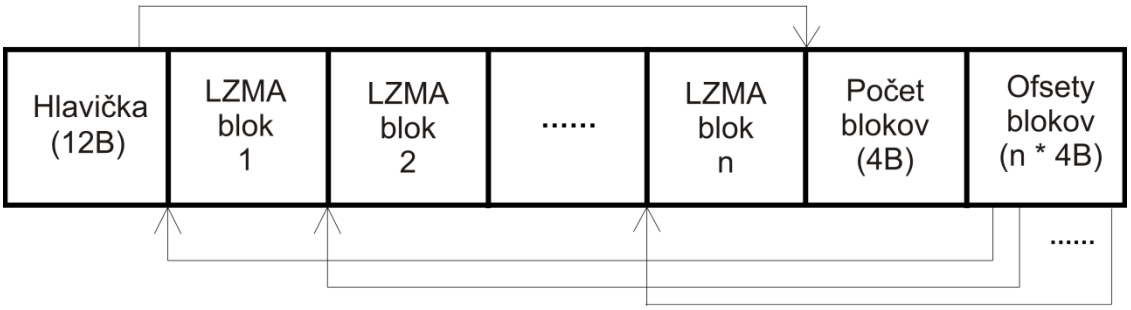


Diagram 4: Kompresný formát dátového archívu

Všetky používané viacbajtové hodnoty sú little endian.

Súbor začína hlavičkou. Štruktúra hlavičky je popísaná v tabuľke 4.

Popis	Veľkosť
„Magické“ číslo (obsahuje hodnotu `0x88022022`)	4
Číslo hlavnej verzie kompresného formátu[1]	2
Číslo vedľajšej verzie kompresného formátu	2
Ofset[2] adresára blokov	4

Tabuľka 4: Štruktúra hlavičky dátového archívu

Ďalej nasleduje viacero blokov skomprimovaných LZMA enkodérom. Podľa dokumentácie priloženej k LZMA SDK [24] má LZMA blok štruktúru, ktorá je popísaná v tabuľke 5.

Popis	Veľkosť
Parametre LZMA	1
Veľkosť slovníka	4
Neskomprimovaná veľkosť	8
Skomprimované dáta	premenlivá

Tabuľka 5: Štruktúra LZMA bloku

Jednotlivé bloky obsahujú jeden alebo viac článkov vo wiki jazyku, doplnených o hlavičku. Táto hlavička obsahuje viacero riadkov oddelených znakom `LF`. Je ukončená prázdnym riadkom. Prvý riadok hlavičky obsahuje názov článku. Toto riešenie bolo zvolené s ohľadom na jednoduchú rozšíriteľnosť. Je relatívne jednoduché pridať podporu iných typov dát, akými sú napríklad obrázky, ktoré bude možné rozlíšiť od wiki článkov pomocou nastaveného MIME typu v hlavičke.

[1] Súčastná implementácia predpokladá verziu 1.0 (tj. hlavná verzia 1 a vedľajšia 0).
[2] Jedná sa o ofset od začiatku súboru.

Za LZMA blokmi nasleduje adresár blokov. Prvé 4 bajty adresára blokov tvorí číslo reprezentujúce počet blokov v súbore. Za ním nasledujú 4-bajtové čísla reprezentujúce ofsety jednotlivých blokov.

Kvôli obmedzeniu súborovému systému FAT32 na veľkosť súborov bolo potrebné v programe implementovať podporu viaczväzkových archívov. Jednotlivé zväzky takéhoto archívu majú rovnaký formát ako jednozväzkový archív. Zmeny sa prejavujú len v súbore s indexom názvov článkov.

7.1.3 Index názvov článkov

Súbor s indexom názvom článkov predstavuje inštanciu index-sekvenčného súboru popísaného v [25]. Jeho formát je znázornený na diagrame 5.

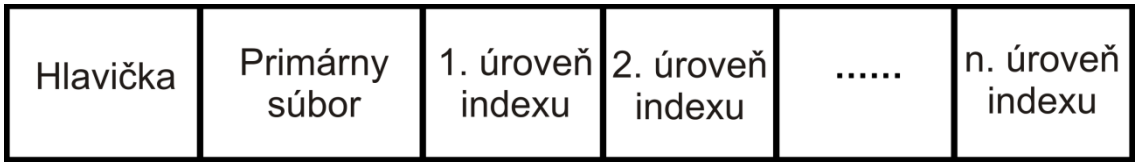


Diagram 5: Formát indexu názvov článkov

Záznamy v tomto súbore sú lexikograficky usporiadané podľa názvu článkov.

Jednotlivé úrovne indexu sú rozdelené do logických blokov určitej veľkosti[1]. Množina prvých položiek každého bloku na určitej úrovni tvorí obsah vyššej úrovne indexu.

Súbor začína hlavičkou. Štruktúra hlavičky je popísaná v tabuľke 6.

Popis	Veľkosť
Príznak viaczväzkového archívu (obsahuje hodnotu 1, ak indexujeme viaczväzkový archív, inak obsahuje hodnotu 0)	1
Počet úrovní indexu	1
Ofset n-tej úrovne indexu (big endian bez znamienka)	4

Tabuľka 6: Štruktúra hlavičky indexu názvov článkov

Ďalej nasleduje primárny súbor (úroveň indexu 0). Štruktúra záznamov, ktoré obsahuje, je popísaná v tabuľke 7. Primárny súbor končí záznamom s názvom „`##BLOCKEND##`“.

Popis	Veľkosť
Názov článku v kódovaní UTF-8 začínajúci a končiaci znakom 0	premenlivá

[1] Jedná sa len o približnú veľkosť. V programe WikiConvert je to implementované tak, že ak je blok väčší alebo rovný ako nastavená limitná hodnota, tak sa vytvorí nový blok.

Číslo zväzku archívu s článkom (ak nie je súbor so skomprimovanými dátami viaczväzkový, tak sa táto položka v indexovom súbore nenachádza)	1
Číslo bloku archívu s článkom (big endian bez znamienka).	4
Ofset začiatku článku v rámci bloku (big endian bez znamienka).	4
Ofset konca článku v rámci bloku (big endian bez znamienka).	4

Tabuľka 7: Štruktúra záznamu v primárnom súbore indexu názvov článkov

Časť s názvom *x. úroveň indexu* reprezentuje x-tú úroveň indexu nad vyššie spomínaným primárnym súborom. Štruktúra záznamov týchto častí je popísaná v tabuľke 8. Každá úroveň indexu končí záznamom s názvom „`##BLOCKEND##`“.

Popis	Veľkosť
Názov článku v kódovaní UTF-8, ktorý sa nachádza na začiatku indexovaného bloku, končiaci znakom 0	premenlivá
Ofset začiatku indexovaného bloku v ďalšej úrovni indexu	4

Tabuľka 8: Štruktúra záznamu v indexe názvov článkov

7.2 WikiConvert - užívateľská dokumentácia

7.2.1 Systémové požiadavky

Pre beh programu WikiConvert je potrebné mať nainštalované behové prostredie Javy (Java Runtime Environment – JRE) a minimálne 64 MB pamäte RAM[1]. Program bol testovaný s JRE vo verzii 1.6. JRE je možné stiahnuť tu [26].

7.2.2 Používanie programu

Program spustíme z príkazového riadku pomocou príkazu:

```
java –jar WikiConvert.jar [<prepínače>][2] <vstupný súbor>
```

Prepínače programu sú popísané v tabuľke 9.

Prepínač	Popis
`-zip <reťazec>`	Nastavenie príkazu pre kompresiu. V programe je k tomuto príkazu pripojený názov výstupného súboru a názvy vstupných súborov. Tento prepínač bol pridaný kvôli testovaniu. (implicitná hodnota: „`wzip –a –d 16`“)
`-names`	Zobrazí menné priestory vstupného súboru a k nim odpovedajúce číselné identifikátory. Následne program skončí.
`-inc <číslo>`	Pridá menný priestor s identifikátorom `číslo` do zoznamu povolených menných priestorov.

[1] Táto veľkosť platí v prípade nenastaveného prepínača „`–links`“. V prípade jeho nastavenia je spotreba pamäte znateľne vyššia – v pamäti je udržiavaná štruktúra, ktorá umožňuje jednoduché zistenie či sa článok s daným názvom nachádza v dátovom súbore.
[2] [] znamená nepovinný parameter.

`-index`	Vytvorí index názvov článkov z už vytvorených nespracovaných dát indexu. Následne program skončí.
`-nolang`	Vymaže z dát článkov odkazy na články v iných jazykoch.
`-links`	Spracuje všetky odkazy v hlavnom mennom priestore a k tým, ktoré sa nenachádzajú v dátovom súbore pridá prefix „`@:`“. Táto funkcia vyžaduje prístup k názvom článkov, ktoré sa nachádzajú v dátovom súbore. Kvôli rýchlosti sa tieto názvy uchovávajú v pamäti, čo znamená vysokú spotrebu pamäte. Java Virtual Machine má zvyčajne implicitne nastavenú maximálnu veľkosť dostupnej pamäte na nejakú fixnú hodnotu. Preto je potrebné nastaviť pomocou parametru JVM, „`-Xmx<veľkosť>`“ vyššiu maximálnu veľkosť dostupnej pamäte pre Javu.
`-b <číslo>`	Nastaví limit veľkosti zväzku archívu. Po dosiahnutí tohto limitu je vytvorený nový archív. Tj. archív článkov sa stáva viaczväzkový. (implicitná hodnota: 1024*1024*1024)
`-f <číslo>`	Nastaví limit počtu dočasných súborov na jednu kompresiu (na jedno spustenie kompresného programu). (implicitná hodnota: 100)
`-s <číslo>`	Nastaví limit veľkosti komprimovaného bloku v bajtoch. Po prevýšení tejto hodnoty je blok uzavretý a je otvorený nový (vždy je však v bloku uložený celý článok). (implicitná hodnota: 100*1024)
`-i <číslo>`	Nastaví limit veľkosti bloku indexu v bajtoch. (implicitná hodnota: 20*1024)

Tabuľka 9: Popis prepínačov programu WikiConvert

Počas behu programu je v konzole zobrazovaný priebeh konverzie.

7.2.3 Odporúčané nastavenia prepínačov programu

Pre správne zobrazenie článkov encyklopédie Wikipédia je nevyhnutné zahrnúť menné priestory s identifikátormi 0 (hlavný menný priestor) a 10 (šablóny - menný priestor „*Template*“ v anglickej Wikipédii). Ďalšími priestormi vhodnými k zahrnutiu sú napríklad priestory s identifikátormi 14 (kategórie - menný priestor „*Category*“ v anglickej Wikipédii) a 100 (portály – menný priestor „*Portal*“ v anglickej Wikipédii).

Limit veľkosti komprimovaného bloku je vhodné zvoliť čo najmenší, pretože pri zobrazovaní článku sa blok rozbaľuje sekvenčne, až pokiaľ sa nerozbalia všetky dáta článku. Pre zistenie optimálnej veľkosti bloku bol vykonaný test konverzie databázového exportu „*simplewiki-20090912-pages-articles.xml*“ s rôznymi nastaveniami limitu veľkosti bloku. Pri teste neboli nastavené prepínače „`-nolang`“ a „`-links`“ Test bol vykonaný na počítači s procesorom Core2Duo na frekvencii 2 GHz. Namerané výsledky testu sú v tabuľke 10. Najvhodnejším kompromisom je veľkosť bloku v rozsahu 20 až 100 KB.

Limit veľkosti bloku	Čas konverzie[1]	Kompresný pomer[2]
1Kb	25:19	30,23%
10Kb	7:19	24,86%
20Kb	5:14	23,43%
50Kb	4:43	21,85%
100Kb	3:45	20,83%
200Kb	3:40	19,93%
500Kb	3:28	18,88%

Tabuľka 10: Výsledky testu limitu veľkosti bloku

Dátové súbory používané pri testovaní programu WikiReader boli vytvorené s použitím prepínačov uvedených v tabuľke 11.

```
-nolang –links -inc 0 –inc 10 –s 51200 –f 100 –i 4096
```

Tabuľka 11: Odporúčané nastavenia prepínačov programu WikiConvert

7.2.4 Testy konverzie

Konverzia databázových exportov bola testovaná na rýchlosť a veľkosť výsledných súborov. Testy boli vykonané na vybraných jazykových mutáciách encyklopédie Wikipédia. Konverzia prebiehala na počítači s procesorom Core2Duo na frekvencii 2 Ghz. Prepínače programu WikiConvert boli nastavené tak, ako sú uvedené v tabuľke 11. Výsledky testov sú v tabuľke 12.

Jazyková mutácia	Veľkosť databázového exportu	Čas konverzie[3]	Veľkosť vytvoreného dátového súboru	Kompresný pomer[2]	Spotreba pamäti
Slovenská	408 MB	0:05:45	82 MB	20%	70 MB
Česká	1,01 GB	0:11:12	208 MB	20%	85 MB
Zjednodušená Anglická	209 MB	0:02:44	36 MB	17%	55 MB
Anglická	24,8 GB	2:48:24	4,98 GB	21%	1.4 GB

Tabuľka 12: Testy konverzie encyklopédie Wikipédia

7.3 WikiConvert – implementácia

7.3.1 Použité technológie

Na program WikiConvert bola kladená požiadavka, aby bol čo najmenej závislý na použitom operačnom systéme. Výber programovacieho jazyka sme obmedzili na skriptovací jazyk, akým je napríklad Perl alebo Python, a jazyk Java. Výsledkom výberu bol

[1] Čas je vo formáte minúty:sekundy.
[2] Pomer veľkosti výsledného dátového súboru k veľkosti pôvodného databázového exportu.
[3] Čas je vo formáte hodiny:minúty:sekundy.

nakoniec jazyk Java, kvôli subjektívne príjemnejšiemu prostrediu a vyššej rýchlosti, ako v prípade skriptovacieho jazyka.

Pre potreby testovania rôznych kompresných metód bola potrebná možnosť zmeny použitej kompresnej metódy. Zvolili sme riešenie, v ktorom kompresiu poskytuje externý kompresný program. Výhodou tohto riešenia je práve vyššie spomenutá rýchla zameniteľnosť kompresnej metódy a jednoduchšia implementácia. Nevýhodou je, že je toto riešenie pomalšie[1].

7.3.2 Štruktúra zdrojových súborov

Zdrojový kód programu WikiConvert obsahuje nasledujúce súbory/triedy:

- `Main.java` – Vstupný bod aplikácie. Vyhodnotí argumenty z príkazového riadku, spustí konverziu a vytvorenie indexov. Nakoniec vytvorí konfiguračný súbor.
- `XMLRead.java` – Obsahuje triedu, ktorá vykonáva vlastnú konverziu databázového exportu. Ku kompresii využíva inštanciu triedy `Compress`. Taktiež vytvorí nespracovaný index názvov článkov. Pojem nespracovaný znamená, že je ešte neusporiadaný a v textovom formáte.
- `IndexSort.java` – Obsahuje triedu, ktorej úlohou je usporiadať nespracovaný index názvov článkov. Výsledkom je usporiadaný súbor v textovom formáte.
- `IndexCreate.java` – Obsahuje triedu, ktorej úlohou je vytvoriť index názvov článkov z usporiadaného indexu v textovom formáte.
- `Compress.java` – Obsahuje triedu, ktorej úlohou je spustiť externý program, ktorý skomprimuje vstupné súbory.

7.3.3 Postup konverzie

Zjednodušene funguje program WikiConvert nasledovne:

1. Vyhodnotenie argumentov príkazovej riadky (trieda `Main`).

[1] Najskôr sú súbory zapísané na disk, potom sú komprimované externým programom a nakoniec sú tieto (dočasné) súbory vymazané.

2. Získanie informácií o projekte z databázového exportu (trieda `XMLRead`).
3. Vlastná konverzia databázového exportu (trieda `XMLRead`).
4. Usporiadanie nespracovaného indexu názvov článkov (trieda `IndexSort`).
5. Vytvorenie finálneho indexu názvov článkov (trieda `IndexCreate`).
6. Uloženie konfiguračného súboru projektu (trieda `Main`).
7. Vymazanie dočasných súborov (trieda `Main`).

7.3.4 Princíp spracovania XML súboru

Vzhľadom na jednoduchú štruktúru databázových exportov neboli na ich spracovanie použité nástroje pre spracovávanie XML súborov obsiahnuté v Jave.

Bolo použité nasledujúce riešenie: Súbor je v cykle čítaný po jednotlivých znakoch. Ak program narazí na znak „<“, prečíta nasledujúci tag a podľa neho a obsahu stavových premenných sa rozhodne čo urobí – napríklad zmení stavovú premennú alebo uloží znak do pamäte. Stavové premenné sú používané pre indikáciu toho, v ktorom elemente sa práve nachádzame (napríklad žiadny element, článok alebo názov článku).

7.3.5 Vlastná konverzia databázového exportu

Vlastnú konverziu databázového exportu tvorí cyklus, ktorý vykonáva tieto kroky:

1. Získanie názvu článku (uložený v tagu `<title>`). Ak názov článku patrí do menného priestoru, ktorý je povolený, tak sa získa obsah článku (uložený v tagu `<text>`). Inak sa číta XML súbor, až pokým sa nenarazí na ďalší tag s názvom článku.
2. Uloženie názvu a obsahu článku do vyrovnávacej pamäte aktuálneho bloku.
3. Pridanie informácie o článku do nespracovaného indexu. Jedna položka nespracovaného indexu obsahuje: názov článku, číslo zväzku archívu, číslo bloku, ofset začiatku článku v bloku a ofset konca článku v bloku.
4. Ak je veľkosť aktuálneho bloku väčšia ako limit veľkosti bloku, vytvorí sa dočasný súbor s obsahom aktuálneho bloku, zväčší sa číslo aktuálneho bloku a vymaže sa obsah vyrovnávacej pamäte aktuálneho bloku.

5. Ak je počet vytvorených dočasných súborov väčší ako limit počtu súborov na jednu kompresiu, tak sa vykoná nasledujúce:

    a. Spustenie externého kompresného programu nastaveným príkazom pre kompresiu (možné špecifikovať cez príkazovú riadku, implicitne má hodnotu: „`wzip a -d 16`“[1]), ku ktorému je pripojený názov aktuálneho archívu a názvy súborov určených na kompresiu.

    b. Ak je veľkosť aktuálneho zväzku archívu väčšia ako limit veľkosti zväzku, tak sa vytvorí nový zväzok. Taktiež sa zväčší číslo zväzku a číslo bloku sa nastaví na 0.

    c. Vymazanie dočasných súborov.

6. Po prečítaní vstupného súboru sa uloží posledný blok a zakomprimujú sa existujúce dočasné súbory.

Pre zmenšenie veľkosti dátového archívu sú v ňom uložené iba názov a obsah článku. Ostatné informácie, ktoré obsahuje databázový export, nie sú ukladané do výsledného súboru.

V prípade, že je nastavený prepínač „`-nolang`“ tak sú pri spracovávaní článku odstránené wiki odkazy na články v iných jazykoch. Tieto odkazy sú špeciálne tým, že začínajú jazykovým prefixom, napríklad „`en:`“, „`sk:`“, „`cs:`“.

V prípade, že je nastavený prepínač „`-links`“, tak je ešte pred začiatkom vlastnej konverzie databázového exportu vytvorená množina názvov článkov, ktoré tento export obsahuje. Táto množina je udržiavaná v pamäti kvôli rýchlemu prístupu k nej. Z toho sa odvíja aj vysoká spotreba pamäti, ktorá dosahuje pri konverzii anglickej Wikipédie až 1.4 GB. Po nastavení tohto prepínača je pri spracovávaní článku pridaný k wiki odkazom v hlavnom mennom priestore, ktoré sa nenachádzajú v dátovom súbore, menný prefix „`@:`“.

7.3.6 Usporiadanie indexu

Trieda `IndexSort` predstavuje implementáciu n-cestného algoritmu vonkajšieho triedenia. Kód triedy je založený na implementácii z [27].

[1] Parameter „`-d 16`“ nastaví veľkosť slovníku používaného v algoritme LZMA na 2^{16} bajtov.

Trieda triedi podľa názvov článkov. Pri porovnávaní však nie sú používané reťazce Javy (`java.lang.String`), ktoré sú v kódovaní UTF-16, ale ich bajtové reprezentácie v kódovaní UTF-8, ktorých číselné hodnoty sú porovnávané. Toto riešenie bolo zvolené s ohľadom na program WikiReader, v ktorom nie sú pri vyhľadávaní v indexe používané viacbajtové hodnoty znakov.

7.3.7 Vytvorenie indexu

Vytvorenie indexu názvov článkov prebieha nasledovne:

1. Zápis prázdnej hlavičky (rezervujeme si miesto na začiatku súboru).
2. Vytvorenie primárneho súboru zo súboru obsahujúceho usporiadaný index. Pri vytváraní je udržiavaná informácia o veľkosti aktuálneho bloku indexu. Ak je veľkosť bloku 0, tak sa do pomocného poľa pridá názov aktuálne čítanej položky usporiadaného indexu a jej ofset vo výslednom súbore indexu.
3. Ak je veľkosť bloku väčšia ako limit veľkosti bloku indexu, tak sa nastaví veľkosť bloku na 0.
4. Vytvorenie ďalších úrovní indexu prebieha analogicky ako vytvorenie primárneho súboru. Namiesto dát zo súboru s usporiadaným indexom sa však používajú dáta z pomocného poľa používaného pri vytváraní predchádzajúcej úrovne indexu. Toto pole obsahuje názvy a ofsety položiek, ktoré sa nachádzajú na začiatku každého bloku predchádzajúcej úrovne indexu.
5. Zápis hlavičky.

Jednotlivé úrovne indexu sú zakončené položkou s názvom „`##BLOCKEND##`“.

Problémom bolo, že v Jave neexistujú číselné typy bez znamienka. Toto je v programe vyriešené pomocou metódy, ktorá dokáže skonvertovať číslo na pole bajtov veľkosti n. Výsledné číslo je n-bajtové číslo vo formáte bigendian bez znamienka.

8 WikiReader – užívateľská dokumentácia

V tejto kapitole je popísaná inštalácia a spôsob používania programu WikiReader.

8.1 Popis programu

Program WikiReader predstavuje jednoduchý webový server. K tomuto serveru sa je možné pripojiť pomocou webového prehliadača.

8.2 Systémové požiadavky

Pre spustenie programu je potrebné mobilné zariadenie s operačným systémom Windows Mobile minimálne vo verzii 5.0 s aspoň 2MB voľnej pamäte RAM[1] a dostatok voľnej pamäte na úložnom zariadení pre uloženie dátových súborov.

8.3 Inštalácia programu

Distribúcia programu neobsahuje inštalačný súbor. Inštaláciu je potrebné previesť manuálne skopírovaním adresára s programom kdekoľvek do zariadenia. Dátové súbory je vhodné, kvôli ich veľkosti, skopírovať na pamäťovú kartu. Následne je potrebné pridať špecifikácie týchto dátových súborov do konfiguračného súboru, ktorý sa nachádza v adresári s programom.

8.4 Spustenie programu

Program je možné spustiť pomocou spustiteľného súboru `WikiReader.exe`.

Po spustení webového prehliadača je možné prehliadať obsah špecifikovaním adresy „*http://localhost*“ s portom, ktorý je nastavený v konfiguračnom súbore.

Použiť je možné prehliadač „Internet Explorer Mobile“. Ostatné prehliadače neumožňujú, kvôli obmedzeniu systému Windows Mobile, prístup k lokálnemu serveru ak neexistuje žiadne aktívne internetové spojenie, napríklad wifi alebo ActiveSync spojenie.

[1] Pamäť potrebná pre webový prehliadač nie je zahrnutá v tomto množstve.

8.5 Prehliadanie obsahu

Po pripojení k serveru sa zobrazí stránka s informáciami o dostupných dátových súboroch. Táto stránka taktiež umožňuje vyhľadávať názvy článkov v jednotlivých dátových súboroch a ukončiť program WikiReader.

Pri vyhľadávaní článkov sa vyhľadáva na čiastočnú zhodu od začiatku názvu článku. Výsledkom vyhľadávania je zoznam článkov, ktorých názvy sú abecedne radené vyššie alebo rovno ako hľadaný článok. V tomto zozname sa je možné posúvať pomocou tlačidiel umiestnených na jeho začiatku a konci.

Pri prehliadaní článku sa na vrchole stránky nachádzajú ovládacie prvky, umožňujúce prechod na začiatočnú stránku s informáciami o dátových súboroch a vyhľadávanie v aktuálnom dátovom súbore.

9 WikiReader – implementácia

V tejto kapitole je popísaný spôsob implementácie programu WikiReader.

9.1 Platforma

Program WikiReader a jeho zásuvné moduly sú vyvíjané pre platformu Windows Mobile. Programovacím jazykom je jazyk C++.

Vývoj pre platformu Windows Mobile je veľmi podobný vývoju pre klasické Windows. Je preň charakteristické:

- Windows Mobile podporuje podmnožinu aplikačného rozhrania Win32 API. Programátor so skúsenosťami s vývojom pre Windows sa vie veľmi rýchlo naučiť programovať pre Windows Mobile.
- Niektoré štandardné funkcie jazyka C a C++ nie sú podporované – chýba napríklad `errno.h`, implementácia funkcií pre prácu s časom a viacero iných funkcií.
- Chýba koncept aktuálneho adresára. Všetky cesty k súborom je treba zadávať v úplnom tvare (od koreňového adresára).
- Systém Windows Mobile vynucuje používanie kódovania Unicode. Ak vo Win32 API na Windows existujú varianty funkcie akceptujúce reťazec ako parameter v kódovaní Unicode a v klasickom ASCII, tak vo Windows Mobile je podporovaná len Unicode verzia tejto funkcie.
- Príkazová riadka nie je podporovaná (tj. výstup z `printf` a podobných funkcií sa navonok neprejaví).
- Pri programovaní pre mobilné zariadenie je treba myslieť na obmedzenia hardvéru. Jedná sa hlavne o menšiu dostupnú pamäť a výkon. Taktiež je treba myslieť na výdrž batérie.

9.2 Základná terminológia

V tejto kapitole je popísaná základná terminológia, ktorá je používaná v popise implementácie programu WikiReader.

Modul

Modulom rozumieme samostatný skompilovaný súbor. Konkrétne na platforme Windows Mobile sa jedná o spustiteľné súbory a dynamicky linkované knižnice.

Zásuvný modul

Pod pojmom zásuvný modul je možné rozumieť dynamicky linkovanú knižnicu. Zásuvné moduly umožňujú rozšíriť funkcie programu bez zmeny hlavného programu. Navyše ponúkajú užívateľom možnosť voľby funkcií, ktoré potrebujú.

Služba

Zásuvný modul môže obsahovať viacero tried, z ktorých každá implementuje jedno z rozhraní popísaných v kapitole 9.7. Pre potreby tejto práce budeme nazývať tieto triedy službami[1].

9.3 Popis práce programu

Základný princíp práce programu je znázornený na diagrame 6.

Program po svojom štarte načíta svoje nastavenia z konfiguračného súboru volaním metódy `initPreferences` triedy `RequestHandler` (1). Nastavenia obsahujú špecifikácie služieb zásuvných modulov. Každá špecifikácia služby obsahuje názvov služby, cestu k súboru zásuvného modulu a identifikátor služby v rámci zásuvného modulu. Názvy služieb môžu byť použité v tomto konfiguračnom súbore ako aj v konfiguračných súboroch dátových súborov na výber požadovanej služby v konkrétnom kontexte. Špecifikované služby sú zaregistrované v manažérovi zásuvných modulov (2). Ďalej nastavenia obsahujú špecifikácie dátových súborov. Každá špecifikácia dátového súboru obsahuje okrem cesty k tomuto súboru[2] aj názov virtuálneho adresára, pod ktorým budú dáta z tohto súboru prístupné užívateľovi, a názov

[1] Sú nazvané službami preto, lebo poskytujú nejakú funkcionalitu ďalším častiam programu (tj. slúžia ďalším častiam programu).
[2] V prípade dátových súborov popísaných v kapitole 7.1 je uvedená cesta ku konfiguračnému súboru daného dátového súboru.

služby typu súborový formát, ktorá slúži k prístupu k tomuto dátovému súboru. Pre každý dátový súbor je vytvorená inštancia tejto služby. Volaním metódy `open` tejto služby je následne otvorený dátový súbor (3).

Nakoniec sa spustí webový server, ktorý pasívne čaká na HTTP požiadavky (4). Po príchode požiadavky je táto požiadavka predaná metóde `handleRequest` triedy `RequestHandler`. Ak sa jedná o požiadavku na ukončenie programu tak sa program ukončí (5). Inak sa podľa adresy požadovaného zdroja rozhodne, ktorej inštancii služby typu súborový formát je pridelené spracovanie tejto požiadavky (6). Postup spracovania požiadavky po predaní službe typu súborový formát je závislé na implementácii tejto služby (7).

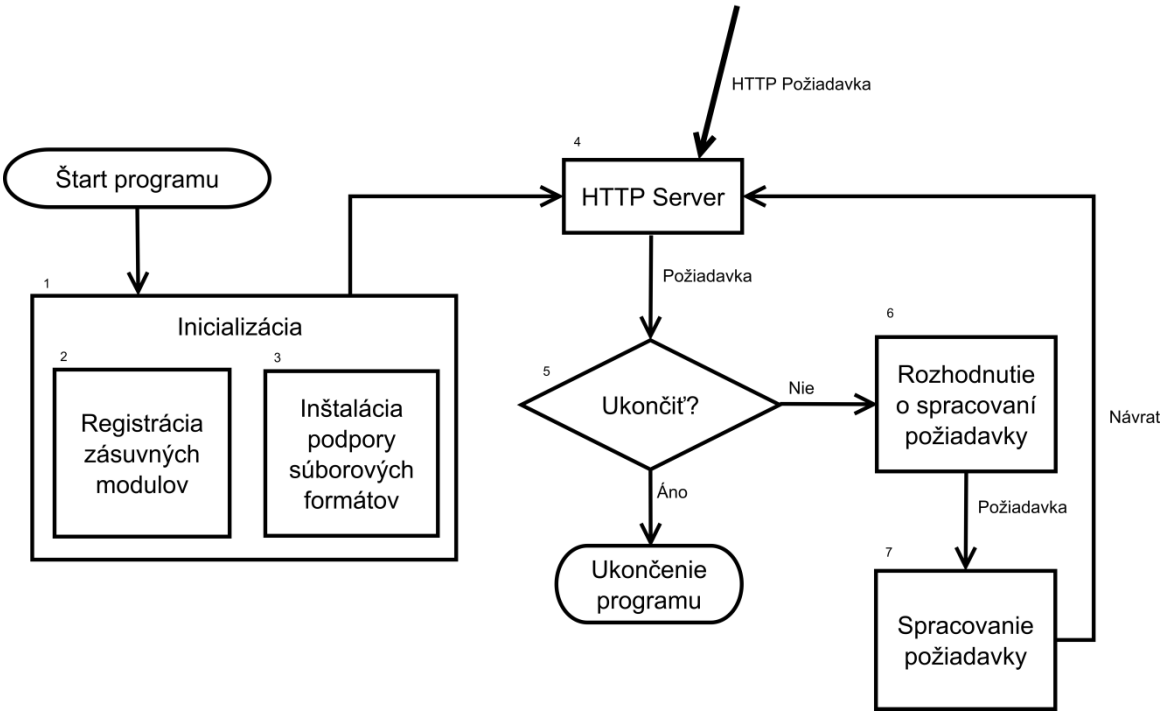


Diagram 6: Hrubá schéma práce programu WikiReader

Na inicializáciu programu a rozhodnutie o spracovaní požiadavky slúži trieda `RequestHandler`. Jedná sa o triedu typu singleton[1]. Inštanciu tejto triedy je možné získať pomocou jej statickej metódy `instance()`.

[1] V programe existuje iba jedna inštancia tejto triedy.

9.4 Dátové štruktúry

Pre výmenu dát medzi rôznymi časťami programu WikiReader je používaných viacero dátových štruktúr.

Požiadavka

Trieda `Request` reprezentuje požiadavku na spracovanie. Táto trieda obsahuje virtuálne metódy na nastavenie a získanie identifikátora požadovaného zdroja (`setId` a `getId`), na zápis čiastočných výsledkov spracovania (`writeChunk`) a na nastavenie alebo získanie parametrov požiadavku (`setParam` a `getParam`). Od tejto triedy dedí trieda `HttpRequest`, ktorá reprezentuje požiadavku webového serveru. Metóda `writeChunk` slúži na zápis do výstupného prúdu. V potomkoch triedy `Request` môže byť implementovaná napríklad zápisom do súboru, pamäte, reťazca alebo soketu.

Adresa v dátovom archíve

Poloha zdroja v dátovom archíve je reprezentovaná triedou `Address`. Táto trieda obsahuje reťazec identifikujúci zdroj[1], zväzok archívu, číslo bloku a ofset začiatku a konca dát požadovaného zdroja v rámci bloku.

Táto trieda bola navrhovaná pre potreby uchovania adresy článku v dátovom archíve s formátom popísaným v kapitole 7.1.2. Je však možné jej použitie aj na uchovanie adresy v iných typoch archívov – napríklad v zip archívoch je možné využiť reťazec identifikujúci zdroj pre uloženie cesty k súboru nachádzajúcom sa v archíve.

Pomocné pole bajtov

Pre návrat väčšieho množstva dát z metód objektov služieb sa používa trieda `ByteBuffer`, ktorá reprezentuje pole bajtov v pamäti. Spolu s ukazateľom na pole bajtov taktiež obsahuje informáciu o veľkosti tohto poľa. Spôsob alokácie pamäte tohto poľa sa môže líšiť medzi rôznymi službami a zásuvnými modulmi. Pre uvoľnenie tejto pamäte je preto potrebné zavolať metódu `releaseResource` tej služby, ktorá alokovala pamäť tomuto poľu.

[1] V prípade wiki článkov sa jedná o názov článku.

9.5 Konfiguračné súbory

Konfiguračné súbory sú vo formáte XML. Tento formát bol zvolený, pretože sa jednoducho vytvára, je štruktúrovaný a je pre človeka ľahko čitateľný. Konfiguračné súbory používané v programe obsahujú len XML tagy - neobsahujú žiadne atribúty tagov.

Prístup ku konfiguračným súborom poskytuje trieda `Preferences`. Táto trieda interpretuje vstupný XML súbor ako stromovú štruktúru. Prístup ku konkrétnej položke konfiguračného súboru je možný pomocou identifikátoru tejto položky. Tento identifikátor obsahuje názvy jednotlivých „rodičovských" položiek oddelené znakom „/" a názov tejto položky. Neobsahuje názov koreňového elementu XML súboru.

Napríklad máme nasledujúci XML súbor:

```
<?xml version="1.0" encoding="utf-8"?>
<config>
  <polozka1>
    <polozka2>data</polozka2>
  </polozka1>
</config>
```

Položku obsahujúcu hodnotu „`data`" potom reprezentuje identifikátor „`/polozka1/polozka2`".

Trieda `Preferences` zapuzdruje prístup ku konfiguračným súborom tak, aby bolo jednoducho možné zmeniť ich formát a implementáciu tejto triedy bez zmeny ostatných častí programu.

K čítaniu XML súborov sa používa XML parser TinyXML. Jeho zdrojový kód a dokumentáciu je možné nájsť v [28].

Popis konfiguračného súboru programu a konfiguračných súborov dátových súborov je uvedený v prílohe C.

9.6 Webový server

Interakcia programu s užívateľom je riešená pomocou jednoduchého HTTP serveru. Základ jeho implementácie pochádza z [29]. Implementácia tohto serveru využíva niektoré štandardné funkcie jazyka C, ktoré nie sú podporované systémom Windows Mobile. Pre ich implementáciu bolo použité riešenie z [30].

Webový server je možné spustiť vytvorením inštancie triedy `webserver`. Konštruktoru tejto triedy je potrebné dodať ako parameter port, na ktorom bude server bežať. V konštruktore je taktiež možné špecifikovať počet vlákien, ktoré bude server používať.

Server po príchode požiadavky predá túto požiadavku na spracovanie metóde `handleRequest` triedy `RequestHandler`. HTTP požiadavky reprezentuje trieda `HttpRequest`, ktorá dedí od triedy `Request`.

Najväčšou zmenou oproti pôvodnej implementácii HTTP serveru bolo pridanie podpory tzv. „chunked transfer encoding“ (v preklade „dávkové prenosové kódovanie“). Toto kódovanie je vhodné použiť vtedy, keď server chce začať prenos odpovede, ale ešte nepozná jej celú dĺžku. Veľmi vhodné je jeho použitie na zasielanie informácií, ktoré sú generované za behu programu. Jednotlivé kusy odpovede je možné zapísať volaním metódy `writeChunk` triedy `HttpRequest`. Volaním metódy `writeFinalChunk` tejto triedy je zase možné ukončiť dávkový prenos. Viac o tejto funkcii HTTP/1.1 je možné nájsť v [31].

9.7 Zásuvné moduly

V budúcnosti sa plánuje rozšíriť program o ďalšie funkcie. Preto bola implementovaná podpora zásuvných modulov.

Každý zásuvný modul musí exportovať nasledujúce funkcie[1]:

- `const char* **getId**()` – Funkcia, ktorá vracia unikátny identifikátor zásuvného modulu.
- `const char* **getServices**()` – Funkcia, ktorá vracia reťazec obsahujúci zoznam identifikátorov služieb, ktoré daný zásuvný modul poskytuje. Jednotlivé identifikátory sú oddelené znakom „`;`“.
- `ServiceBase* **getInstance**(const char *service)` – Funkcia, ktorá vracia inštanciu objektu služby s identifikátorom `service`. V tomto identifikátore nezáleží na veľkosti písmen.

[1] Pre export bola použitá deklarácia `extern "C" __declspec(dllexport)`.

Zásuvný modul je v programe reprezentovaný triedou `Plugin`. Na prácu so zásuvnými modulmi slúži manažér zásuvných modulov, ktorý je reprezentovaný triedou `PluginManager`. Táto trieda ponúka metódy pre registráciu názvov služieb používaných v programe s konkrétnym zásuvným modulom a jeho službou. Umožňuje taktiež vytváranie inštancií objektov registrovaných služieb. Trieda predstavuje objekt typu singleton. Inštanciu tejto triedy je možné získať pomocou jej statickej metódy `instance()`. Mechanizmus registrácie služieb bol zvolený s ohľadom na jednoduchú editáciu konfiguračných súborov.

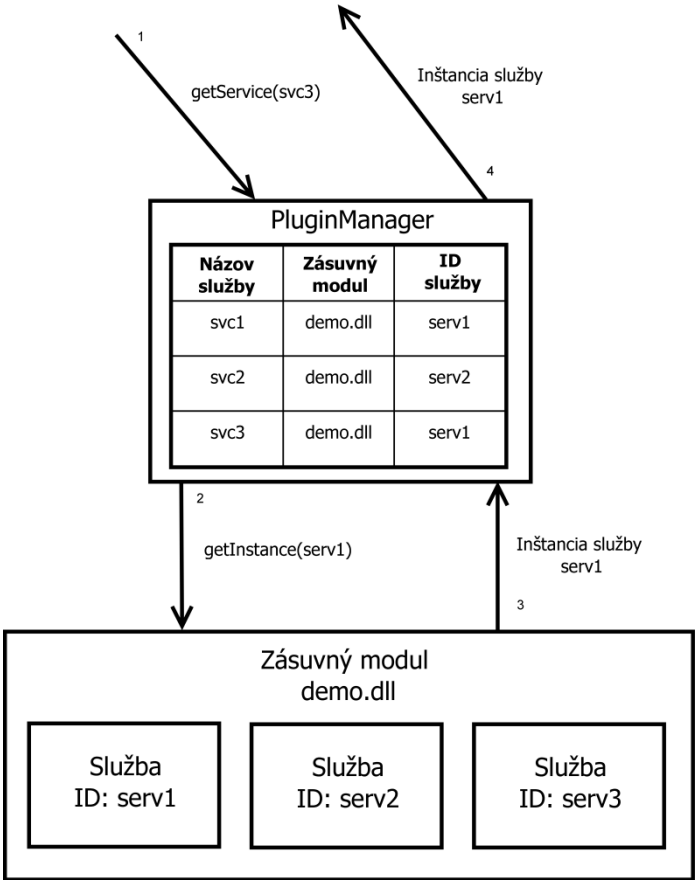


Diagram 7: Schéma práce so zásuvnými modulmi

Na diagrame 7 je znázornený základný princíp práce so zásuvnými modulmi. Manažér zásuvných modulov obsahuje informácie o registrovaných službách (nadefinovaný názov služby, zásuvný modul so službou a identifikátor služby v rámci zásuvného modulu). Ak potrebuje nejaká časť programu inštanciu nejakej služby tak si túto inštanciu vyžiada od manažéra zásuvných modulov, predaním názvu požadovanej služby jeho metóde `getService` (1). Manažér zásuvných modulov zavolá funkciu `getInstance` zásuvného modulu, ktorý bol zaregistrovaný s požadovaným názvom služby. Funkcii

`getInstance` predá identifikátor služby, ktorý bol zaregistrovaný s požadovaným názvom služby (2). Zásuvný modul vytvorí inštanciu požadovanej služby a vráti ukazateľ na túto inštanciu ako návratovú hodnotu funkcie `getInstance` (3). Manažér zásuvných modulov zase tento ukazateľ vráti ako návratovú hodnotu metódy `getService` (4).

V tabuľke 13 sa nachádza popis služieb, ktoré sú štandardne zaregistrované v manažérovi zásuvných modulov (nie je potrebné uviesť ich špecifikácie v konfiguračnom súbore). Registráciu týchto služieb je možné v konfiguračnom súbore zmeniť (špecifikovaním služby s rovnakým názvom).

Názov služby	Zásuvný modul / identifikátor služby	Typ služby
WikiFile	WikiData.dll / WikiFile	Súborový formát
WZip	WikiData.dll / WZip	Dekompresia
WikiTitleIndex	WikiData.dll / WikiTitleIndex	Index
WikiContent	WikiData.dll / WikiContent	Konverzia obsahu
FileSystem	FileSystem.dll / FileSystem	Súborový formát

Tabuľka 13: Štandardné služby v programe WikiReader

Rozhrania služieb sú v jazyku C++ implementované ako abstraktné triedy obsahujúce viacero virtuálnych metód.

Ďalej nasleduje popis typov služieb používaných v programe aj s ich rozhraniami. Ak je v popise rozhrania služby uvedené, že je daná metóda voliteľná, tak ju netreba implementovať, pretože existuje jej implicitná implementácia v rozhraní.

Základné rozhranie služieb

Všetky rozhrania služieb dedia od rozhrania `ServiceBase`. Toto rozhranie obsahuje metódy uvedené v tabuľke 14.

Metóda	Popis
`~ServiceBase()`	Virtuálny deštruktor.
`void release()`	Vymazanie inštancie triedy.
`void releaseResource(void *)`	Vymazanie objektu, vytvoreného touto inštanciou triedy.
`void addService(ServiceBase* plug, const std::string& name)`	Nastavenie služby, ktorú potrebuje táto inštancia služby. Táto metóda je voliteľná. Implicitne nevykoná nič.
`void setPreference(const std::string& name, const std::string& data)`	Nastaví parameter objektu s názvom `name` na hodnotu `data`. Táto metóda je voliteľná. Implicitne nevykoná nič.
`void setPreferenceW(const std::string& name, const`	Nastaví parameter objektu s názvom `name` na hodno-

std::wstring& data)	tu data. Táto metóda je voliteľná. Implicitne nevykoná nič.

Tabuľka 14: Popis metód rozhrania `ServiceBase`

Služba dekompresia

Úlohou služby typu dekompresia je dekomprimovať komprimovaný archív. Každá taká-to služba musí implementovať rozhranie `ServiceDecompress`. Metódy, potrebné pre implementáciu tohto rozhrania, sú uvedené v tabuľke 15.

Metóda	Popis
bool **open**(const std::wstring& archFile)	Otvorenie súboru s archívom. Parameter archFile udáva cestu k súboru archívu. Metóda vracia hodnotu true, ak sa otvorenie podarilo, false inak.
void **close**()	Zatvorenie súboru s archívom.
bool **extract**(size_t block, size_t start, size_t end, ByteBuffer& outdata)	Dekomprimovanie bloku block komprimovaného súboru od ofsetu start k ofsetu end. Výsledok sa ukladá v outdata. Vracia true, ak sa nevyskytla chyba, false inak.
bool **extract**(const std::string& id, size_t start, size_t end, ByteBuffer& outdata)	Dekomprimovanie položky archívu s názvom id od ofsetu start k ofsetu end. Výsledok sa ukladá v outdata. Táto metóda je voliteľná. Implicitne vracia hodnotu false.

Tabuľka 15: Popis metód rozhrania `ServiceDecompress`

Služba index

Úlohou služby typu index je získať adresu položky v komprimovanom archíve. Každá takáto služba musí implementovať rozhranie `ServiceIndex`. Metódy, potrebné pre implementáciu tohto rozhrania, sú uvedené v tabuľke 16.

Metóda	Popis
bool **open**(const std::wstring& indexFile)	Otvorenie súboru s indexom. Parameter indexFile udáva cestu k indexovému súboru. Metóda vracia hodnotu true, ak sa otvorenie podarilo, false inak.
void **close**()	Zatvorenie súboru s indexom.
Address **exactMatch**(Request& req)	Dotaz na úplnú zhodu s požiadavkou req. Vracia adresu odpovedajúceho zdroja.
void **partialMatch**(Request& req, std::vector<Address>& out)	Dotaz na čiastočnú zhodu s požiadavkou req. Výsledné adresy sa uložia do vektora out.

Tabuľka 16: Popis metód rozhrania `ServiceIndex`

Služba súborový formát

Úlohou služby typu súborový formát je poskytovať prístup k dátovým súborom. Každá takáto služba musí implementovať rozhranie `ServiceFileFormat`. Metódy, potrebné pre implementáciu tohto rozhrania, sú uvedené v tabuľke 17.

Metóda	Popis
`bool` **`open`**`(const std::wstring& dataFile)`	Otvorenie dátového súboru. Parameter `dataFile` udáva cestu k súboru. Metóda vracia hodnotu `true`, ak sa otvorenie podarilo, `false` inak.
`void` **`close`**`()`	Zatvorenie súboru.
`bool` **`process`**`(Request& req)`	Spracovanie požiadavky `req`. Metóda vracia hodnotu `false` ak nastala počas spracovania chyba.
`bool` **`getData`**`(const std::string& id, ByteBuffer& outdata)`	Získanie dát s identifikátorom `id` z dátového súboru. Výsledok sa ukladá v `outdata`. Metóda vracia hodnotu `false`, ak nastala počas spracovania chyba. Táto metóda je voliteľná. Implicitne vracia hodnotu `false`.
`Preferences*` **`getPreferences`**`()`	Získanie ukazateľa na objekt typu `Preferences`, ktorý ponúka prístup k súboru s konfiguráciou dátového súboru. Táto metóda je voliteľná. Implicitne vracia hodnotu `NULL`.
`std::string` **`getInfo`**`()`	Získanie informácií o dátovom súbore.

Tabuľka 17: Popis metód rozhrania `ServiceFileFormat`

Služba konverzia obsahu

Úlohou služby typu konverzia obsahu je konverzia dát. Využitím môže byť, napríklad konverzia dát z wiki jazyka, konverzia obrázkov na iný formát alebo konverzia dokumentu pre zobrazenie na mobilnom zariadení. Každá takáto služba musí implementovať rozhranie `ServiceContent`. Metódy, potrebné pre implementáciu tohto rozhrania, sú uvedené v tabuľke 18.

Metóda	Popis
`bool` **`convert`**`(const ByteBuffer& input, Request& req)`	Konverzia vstupu `input`. Výsledok je zapisovaný pomocou metódy `writeChunk` objektu požiadavky `req`. Metóda vracia hodnotu `false`, ak nastala počas spracovania chyba.

Tabuľka 18: Popis metód rozhrania `ServiceContent`

10 WikiReader – zásuvný modul WikiData

V tejto kapitole je popísaná implementácia zásuvného modulu WikiData, ktorý poskytuje služby pre spracovanie požiadavky na článok z dátových súborov popísaných v kapitole 7.1.

10.1 Služby

Zásuvný modul `WikiData.dll` poskytuje služby s nasledujúcimi identifikátormi:

- `WikiFile` – Služba typu súborový formát. Ponúka prístup k dátovým súborom.
- `WZip` – Služba typu dekompresia. Ponúka prístup k jednému zväzku z dátového archívu. Pre každý zväzok dátového archívu je vytváraná samostatná inštancia tejto služby. Implementáciu tejto služby tvorí mierne upravený kód dekompresnej časti programu WZip.
- `WikiTitleIndex` – Služba typu index. Ponúka prístup k indexu názvov článkov.
- `WikiContent` – Služba typu konverzia obsahu. Slúži na konverziu dát z wiki jazyka do HTML. Počas konverzie je výsledok konverzie priebežne odosielaný k zdroju požiadavky.

Vyššie spomenuté služby sú v programe implementované triedami s rovnakými názvami ako identifikátory týchto služieb.

10.2 Popis práce modulu

Dátový súbor je otvorený pri vytvorení inštancie služby `WikiFile` v hlavnom module programu. Otvorenie prebieha volaním metódy `open` tejto služby, ktorej parameter je cesta ku konfiguračnému súboru daného dátového súboru. Metóda `open` vytvorí všetky služby potrebné pre spracovanie požiadavku na článok z daného dátového súboru. Konkrétne sú vytvorené nasledujúce služby:

- služba typu index
- služba typu konverzia obsahu
- služby typu dekompresia – Pre každý zväzok dátového archívu je vytvorená jedna inštancia dekompresnej služby, ktorá umožní rozbalenie dát z tohto zväzku. Toto riešenie bolo zvolené kvôli tomu, aby program pri každej požiadavke na rozbalenie článku neotváral súbor so zväzkom, ktorý obsahuje žiadaný článok.

Názvy konkrétnych služieb sú špecifikované v konfiguračnom súbore. V ďalšom texte predpokladáme, že sú špecifikované názvy práve tých služieb, ktoré sa nachádzajú v module WikiData.

Na diagrame 8 je znázornená základná schéma spracovania článku v module WikiData.

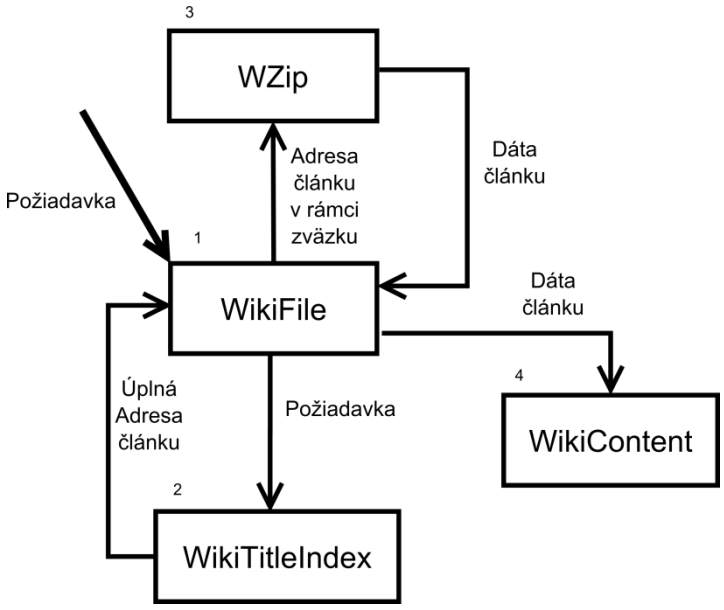


Diagram 8: Schéma spracovania požiadavky na článok z dát Wikipédie

Požiadavka na článok je v hlavnom module programu predaná na spracovanie službe `WikiFile` (1), zavolaním jej metódy `process`. Podľa obsahu požiadavky sa buď vyhodnotí špeciálna požiadavka, ktorou je napríklad vyhľadanie článku v indexe, alebo sa vyhodnotí požiadavka na článok.

Pri vyhodnocovaní požiadavky na článok je potrebné najskôr vyhľadať adresu článku v dátovom archíve. Preto je požiadavka predaná metóde `exactMatch` služby `WikiTitleIndex` (2). Ak požadovaný článok existuje, tak vráti táto metóda adresu článku,

inak vracia neplatnú adresu. V prípade vrátenia neplatnej adresy služba `WikiFile` informuje užívateľa, že článok neexistuje, a ukončí spracovanie.

V prípade platnej adresy, je táto adresa predaná na spracovanie službe `WZip` (3), ktorá reprezentuje zväzok archívu špecifikovaný v adrese, zavolaním jej metódy `extract`. Táto metóda dostane ako parameter inštanciu triedy `ByteBuffer`, v ktorej vráti výsledné dáta článku.

Nakoniec sú dáta článku a požiadavka predané na spracovanie službe `WikiContent` (4), zavolaním jej metódy `convert`. Tieto dáta sú spracované a postupne odosielané pomocou metódy `writeChunk` vstupnej požiadavky.

Kvôli urýchleniu zobrazovania článku a menšej spotrebe pamäti webového prehliadača je článok rozdelený na viacero podstránok. Jednotlivé podstránky obsahujú vždy jednu sekciu oddelenú nadpisom. Odkazy na jednotlivé sekcie článku sa nachádzajú na konci každej čiastočnej stránky článku vo forme obsahu. Dáta článku sú uchovávané v pamäti až pokiaľ program neprijme požiadavku na iný článok.

10.3 Konverzia wiki jazyka do HTML

Konverziu článkov vo wiki značkovacom jazyku do HTML poskytuje služba `WikiContent`. Konverziu je možné spustiť volaním metódy `convert` tejto služby.

Na konverziu bola kladená požiadavka, aby bola čo najrýchlejšia. Urýchlenia spočívali hlavne v minimalizácií počtu alokácií pamäte v priebehu konverzie. Taktiež bolo obmedzené použitie klasických reťazcov jazyka C++. Pri konverzii sú prakticky používané len ako výstupná vyrovnávacia pamäť, ktorá však bola pred začiatkom konverzie predalokovaná.

10.3.1 Dátové štruktúry

Pre ukladanie stavu konverzie slúži trieda `WorkData`. Táto trieda obsahuje stavové premenné, ktoré indikujú stav konverzie – napríklad nachádzame sa v odstavci, v časti s tučným písmom, v tabuľke. Ďalej umožňuje jednoduchý prístup k vnútornému reťazcu, ktorý obsahuje neskonvertované dáta.

Pre jednoduchý prístup k dátam článku uloženým v objekte `ByteBuffer` slúži trieda `PreprocessWorkData`, ktorá ponúka podobné API pre prístup k svojim dátam ako trieda `WorkData`.

Ak je potrebné pri konverzii uložiť v pamäti nejakú časť vstupného reťazca, tak je pre tento účel použitá trieda `stringRef`, ktorá predstavuje referenciu na časť reťazca. Je možné z nej iba čítať.

10.3.2 Postup konverzie

Konverzia z wiki jazyka do HTML prebieha nasledovne:

1. Na začiatku konverzie sú vytvorené inštancie tried `WorkData` a `PreprocessWorkData`. Inštancia `WorkData` je vytvorená ako prázdna, inštancia `PreprocessWorkData` obsahuje nespracované dáta, ktoré dostala služba ako parameter. Taktiež je vytvorená inštancia triedy `string` zo štandardnej knižnice jazyka C++. Táto trieda slúži v programe ako reťazec, do ktorého sa ukladá výsledok konverzie. Volaním metódy `reserve` je predalokovaná dostatočná pamäť pre tento reťazec[1].

2. Predspracovanie určitej časti zo vstupu[2]. Pri predspracovaní je objekt `PreprocessWorkData` prechádzaný po jednotlivých znakoch. V priebehu predspracovania sú nahradené HTML entity odpovedajúcimi znakmi a sú vypustené komentáre. Ostatné znaky sú zapisované na výstup. V neskorších verziách programu sa tu taktiež budú spracovávať šablóny. Výstup predspracovania je ukladaný v objekte `WorkData`.

3. Spracovanie predspracovaného vstupu. Pri spracovaní predspracovaného vstupu je objekt `WorkData` prechádzaný po jednotlivých znakoch. Výstup spracovania je ukladaný do reťazca vytvoreného v 1. kroku. Ak je tento reťazec dlhší ako medzná hodnota, tak sa pomocou metódy `writeChunk` triedy `Request` zapíše časť odpovede a vymaže sa obsah tohto reťazca. Vlastné spracovanie wiki jazyka prebieha formou automatu, ktorý si ukladá svoj stav v stavových premenných objektu `WorkData`.

[1] V súčastnej implementácii sa jedná o 8 KB.
[2] V súčasnej implementácii sa jedná o 10 paragrafov.

4. Kroky 3 a 4 sa opakujú, pokiaľ nie je spracovaný celý vstup.

5. Ukončenie konverzie. Pri ukončovaní konverzie sú ukončené otvorené elementy jazyka HTML – napríklad otvorené tabuľky a odstavce. Taktiež je zapísaná posledná časť odpovede.

10.4 Podporované elementy wiki jazyka

Popis syntaxe wiki jazyka podporovanej systémom MediaWiki je možné nájsť v [32]. Služba WikiContent implementuje vhodnú podmnožinu tohto jazyka. V tejto kapitole sú popísané základné elementy podporovanej syntaxe. V príkladoch je vždy na ľavej strane zdrojový text vo wiki syntaxi a napravo jemu odpovedajúci HTML kód.

10.4.1 Základné formátovanie

Tučné písmo a kurzíva

Text, ktorý má byť zobrazený tučným písmom, je uzavretý do 3 apostrofov. Napríklad:

'''text'''	text

Ak má byť text zobrazený kurzívou, tak je uzavretý do 2 apostrofov. Napríklad:

''text''	text

Tučné písmo a kurzívu možno kombinovať – tj. ak je text uzavretý do 5 apostrofov, tak je zobrazený tučným písmom aj kurzívou. Napríklad:

'''''text'''''	text

Použitím 1 apostrofu je možné zobraziť apostrof. Napríklad:

'text'	'text'

Navyše všetky apostrofy, ktorými nie je možné uplatniť formátovanie tučným písmom alebo kurzívou sú súčasťou výstupu. Napríklad:

''''text''''	'text'

Nadpisy

Text nadpisu je uzavretý medzi znakmi „=“, ktoré sa nachádzajú na začiatku riadku. Ich počet odpovedá úrovni nadpisu. Nadpisy môžu byť 1. až 6. úrovne. Príklad použitia:

===Text===	Text

Horizontálna čiara

Ak sú na začiatku riadku 4 a viac znakov „−“ tak sa vytvorí horizontálna čiara. Napríklad:

Odstavce

Odstavce sú automaticky vytvárané a ukončované, ak je to potrebné. Začať nový odstavec je možné použitím prázdneho riadku. Napríklad:

Text odstavca Nový odstavec	Text odstavca Nový odstavec

Preformátovaný text

Zobrazenie textu bez interpretovania elementov wiki jazyka je možné pomocou tagu <nowiki> alebo <pre>. Oba majú spoločné, že zabraňujú interpretovať wiki značkovací jazyk. Text v <pre> je však zobrazený presne tak, ako je napísaný v zdrojovom kóde – tj. aj s viacerými medzerami a riadkami. Napríklad:

<nowiki>Text s ''wiki''' [[element]]mi</nowiki>	Text s '''wiki''' [[element]]mi
<pre>Text s '''wiki''' [[element]]mi</pre>	<pre>Text s '''wiki''' [[element]]mi</pre>

Ak sa na začiatku riadku nachádza medzera, tak je riadok zobrazený podobne ako pri použití tagu <pre>. Elementy wiki jazyka sú však na tomto riadku interpretované. Napríklad:

Text s '''wiki''' [[element]]mi	<pre>Text s wiki
	elementmi
	</pre>

10.4.2 Odkazy

Wiki značkovací jazyk podporuje viacero druhov odkazov.

Vnútorné odkazy

Vnútorné odkazy fungujú v rámci wiki projektu. Odkaz na článok je možné vytvoriť uzavretím názvu článku do 2 hranatých zátvoriek. Napríklad:

`[[demo]]`	`<a href="demo">demo</a>`

Okrem klasických odkazov sa môže jednať taktiež o obrázky, ktoré sa nachádzajú vo svojom vlastnom mennom priestore. Odkazy môžu obsahovať viacero nepovinných parametrov, ktoré sú oddelené znakom „|".

Klasické odkazy môžu obsahovať parameter, ktorý špecifikuje text, ktorý sa zobrazí namiesto implicitne zobrazovaného názvu odkazovaného článku. Napríklad:

`[[demo|popis]]`	`<a href="demo">popis</a>`

Obrázky môžu obsahovať viacero parametrov – napríklad rozmery obrázku. V programe WikiReader sa však odkazy na obrázky odkazujú na originálnu stránku s obrázkom a podporovaný je len parameter, ktorý reprezentuje popis obrázku.

Externé odkazy

Odkazy na externé zdroje je možné uzavrieť do hranatých zátvoriek. Zobrazovaný text je možné špecifikovať po medzere za odkazom. Ak nie je špecifikovaný zobrazovaný text, tak je odkaz automaticky očíslovaný. Napríklad:

`[http://www.cuni.cz]`	`<a href="http://www.cuni.cz">[1]</a>`
`[http://www.cuni.cz UK Praha]`	`<a href="http://www.cuni.cz">UK Praha</a>`

10.4.3 Zoznamy

Je podporovaná široká škála možností pri definovaní rôznych druhov zoznamov. Jednotlivé zoznamy je navyše možné kombinovať. Zoznamy začínajú jedným alebo viacerými znakmi zo znakov „*", „#", „;" a „:", ktoré sa nachádzajú na začiatku riadku. Počet a typ týchto znakov určuje úroveň zanorenia a typ jednotlivých úrovní zoznamu.

Podporované sú nasledujúce druhy zoznamov:

- nečíslovaný zoznam – Je špecifikovaný znakom „*".
- číslovaný zoznam – Je špecifikovaný znakom „#".

- zoznam definícií – Definovaný termín je špecifikovaný znakom „`;`“, definícia potom znakom „`:`“. Pri použitím samostatného znaku „`:`“ je možné odsadiť text. Znak „`:`“ môže byť navyše použitý aj na riadku s definovaným termínom.

Príklady rôznych druhov zoznamov sa nachádzajú v tabuľke 19.

WIKI	HTML

`* aaa`
`** bbb`
`** ccc`
`* ddd`
	`<ul>`
	`<li>aaa`
	`<ul>`
	`<li>bbb</li>`
	`<li>ccc</li>`
	`</ul>`
	`</li>`
	`<li>ddd</li>`
	`</ul>`

`# Číslovaný zoznam`
`# Pokračovanie úrovne`
`## Ďalšia úroveň`
	`<ol>`
	`<li>Číslovaný zoznam</li>`
	`<li>Pokračovanie úrovne`
	`<ol>`
	`<li>Ďalšia úroveň</li>`
	`</ol>`
	`</li>`
	`</ol>`

`* Nečíslovaný`
`*# Vnorený číslovaný`
`*# Pokračovanie číslovaného`
`*#* Vnorený nečíslovaný`
	`<ul>`
	`<li>Nečíslovaný`
	`<ol>`
	`<li>Vnorený číslovaný</li>`
	`<li>Pokračovanie číslovaného`
	`<ul>`
	`<li>Vnorený nečíslovaný</li>`
	`</ul>`
	`</li>`
	`</ol>`
	`</li>`
	`</ul>`

`: Odsadenie textu`
`:: Väčšie odsadenie`
	`<dl>`
	`<dd>Odsadenie textu`
	`<dl>`
	`<dd>Väčšie odsadenie</dd>`
	`</dl>`
	`</dd>`
	`</dl>`

`Zoznam definícií:`
`;Pojem`
`: definícia`
`: ďalšia definícia`
`; Iný pojem : definícia`
	`Zoznam definícií:`
	`<dl>`
	`<dt>Pojem</dt>`
	`<dd>definícia</dd>`
	`<dd>ďalšia definícia</dd>`
	`<dt>Iný pojem</dt>`
	`<dd>definícia</dd>`
	`</dl>`

Tabuľka 19: Príklady syntaxe zoznamov

10.4.4 Tabuľky

Pre vytvorenie tabuliek vo wiki článkoch slúži špeciálna syntax pre tabuľky, ktorá je prehľadnejšia než tabuľky v HTML.

Tabuľky začínajú znakmi „{|" a končia znakmi „|}". Na riadku začínajúcom znakmi „{|" je možné uviesť HTML atribúty tabuľky. Napríklad:

{| class="wikitable" |}	<table class="wikitable"> </table>

Znaky „|+" nachádzajúce sa na začiatku riadku reprezentujú riadok, ktorý obsahuje nadpis tabuľky. Napríklad:

{| |+ Nadpis tabuľky |}	<table> <caption>Nadpis tabuľky</caption> </table>

Riadky tabuľky je možné vytvoriť pomocou znakov „|-" nachádzajúcich sa na začiatku riadku. Po tomto znaku môžu nasledovať HTML atribúty.

Obsah jednotlivých buniek tabuľky sa nachádza za znakom „|", ktorý sa nachádza na začiatku riadku. Na riadku začínajúcom znakom „|" sa môže vyskytovať viacero buniek. Ich obsah je oddelený znakmi „||". Každá bunka môže obsahovať HTML atribúty. Tieto atribúty je možné uviesť na začiatku bunky a sú ukončené znakom „|". Napríklad:

{| |- style="background:blue" | d || e || f |- | g || style="color:red"| h || i |}	<table> <tr style="background:blue"> <td>d</td><td>e</td><td>f</td> </tr> <tr> <td>g</td> <td style="color:red">h</td> <td>i</td> </tr> </table>

Riadok začínajúci znakom „!" označuje bunku s nadpisom. Ak sa nachádza táto bunka na nejakom riadku tabuľky (za znakom „|-"), tak sa jedná o nadpis riadku. Ak sa nachádza mimo riadku tabuľky, tak sa jedná o nadpis stĺpcov tabuľky. Na riadku začínajúcom znakom „!" sa môže vyskytovať viacero buniek. Ich obsah je oddelený znakmi „!!" alebo „||". Napríklad:

{| ! Nadpis stlpca 1 ! Nadpis stlpca 2 ! Nadpis stlpca 3 |- ! Nadpis riadku | d || f |}	<table> <tr> <th>Nadpis stlpca 1</th> <th>Nadpis stlpca 2</th> <th>Nadpis stlpca 3</th> </tr> <tr> <th>Nadpis riadku</th> <td>data</td> <td>data</td> </tr> </table>

Tabuľka môže obsahovať taktiež vnorené tabuľky. Znaky „{|" označujúce začiatok vnorených tabuliek sa musia vyskytovať na samostatnom riadku.

10.5 Vyhľadávanie v indexe

Vyhľadávanie adresy wiki článku v rámci archívu článkov umožňuje služba `WikiTitleIndex`. Pri vyhľadávaní záleží na veľkosti písmen. Závislosť na veľkosti písmen plynie z toho, že v názvoch článkov wiki projektov záleží na veľkosti písmen, index je ukladaný v kódovaní UTF-8[1] a pri vyhľadávaní v indexe nie sú kvôli rýchlosti používané viacbajtové hodnoty znakov[2].

Vyhľadávať je možné pomocou dotazu na úplnú zhodu použitím metódy `exactMatch` a čiastočnú zhodu (od začiatku názvu článku) použitím metódy `partialMatch`.

[1] Kvôli úspore miesta. UTF-16 a dalsie varianty kodovania Unicode zbytočne plytvajú miestom na disku.
[2] Bolo by potrebné prevádzať pri vyhľadávaní názvy článkov v indexe v kódovaní UTF-8 na UTF-16.

11 WikiReader - prehľad zdrojového kódu a kompilácie

V tejto kapitole je popísaná štruktúra zdrojového kódu programu WikiReader a popis jeho kompilácie.

11.1 Kompilácia programu

Ku kompilácii programu je možné použiť projekty vygenerované prostredím Visual Studio 2008, ktoré sú súčasťou zdrojového kódu programu. Ku kompilácii je treba mať nainštalované Windows Mobile SDK. Program bol testovaný s nasledujúcimi SDK:

- Windows Mobile 6 Professional SDK (zdrojom je [33]).
- Windows Mobile 5.0 SDK (zdrojom je [34]).

Pri kompilácii je potrebné mať zapnuté generovanie behových informácií o typoch (*Run-Time Type Info*)[1] a výnimky jazyka C++.

11.2 Moduly programu

Skompilovaný program je tvorený z viacerých samostatných modulov:

- `WikiReader.exe` – Predstavuje hlavný modul programu.
- `Shared.dll` – Obsahuje súčasti aplikácie, ktoré sú zdieľané medzi jednotlivými modulmi. Tento modul je staticky linkovaný s ostatnými modulmi.
- `WikiData.dll` - Predstavuje zásuvný modul, ktorý poskytuje služby pre spracovanie požiadavku na článok z offline dát Wikipédie.
- `FileSystem.dll` – Predstavuje zásuvný modul, ktorý obsahuje službu typu súborový formát, ktorá sprístupňuje adresár súborového systému užívateľovi prostredníctvom HTTP serveru. Tento modul je používaný pre sprostredkovanie CSS súboru so štýlom.

[1] Je to vyžadované pretože v programe sú používané volania `dynamic_cast`, ktoré túto funkciu jazyka C++ vyžadujú.

11.3 Štruktúra zdrojového kódu

Súbory zdrojového kódu programu sú organizované do viacerých adresárov. Jednotlivé adresáre odpovedajú modulom programu. Súbory, ktoré nie sú uvedené v tomto prehľade, nie sú dôležité pre samotné pochopenie implementácie programu. Adresáre sú uvádzané relatívne k adresáru so zdrojovým kódom.

Projekty vygenerované prostredím Visual Studio nie sú uvedené v tomto prehľade. Projekty jednotlivých modulov sa nachádzajú v adresároch so zdrojovým kódom týchto modulov. Projekt, ktorý obsahuje všetky moduly programu, sa nachádza v adresári so zdrojovým kódom.

11.3.1 Adresár `WikiReader`

V adresári `WikiReader` sa nachádza zdrojový kód hlavného modulu - `WikiReader.exe`. Obsahuje nasledujúce súbory:

- `RequestHandler.h/.cpp` – Obsahuje triedu `RequestHandler`.
- `main.cpp` – Vstupný bod aplikácie.

Adresár `WikiReader` obsahuje taktiež adresár `HTTPServer` so zdrojovým kódom HTTP serveru. Tento adresár obsahuje nasledujúce súbory.

- `HttpRequest.h` – Obsahuje triedu `HttpRequest`.
- `Socket.h/.cpp` – Obsahuje triedy reprezentujúce TCP soket a TCP server.
- `UrlHelper.h/.cpp` - Obsahuje pomocné funkcie pre prácu s URL adresami a požiadavkou klienta.
- `webserver.h/.cpp` - Obsahuje triedu `webserver`.

11.3.2 Adresár `Shared`

V adresári `Shared` sa nachádza zdrojový kód modulu `Shared.dll`. Obsahuje nasledujúce súbory:

- `Plugin.h/.cpp` – Obsahuje triedy `Plugin` a `PluginManager`.
- `Preferences.h/.cpp` – Obsahuje triedu `Preferences`.

- `Request.h` – Obsahuje triedu `Request`.
- `StringUtils.h/.cpp` – Obsahuje pomocné triedy pre prácu s reťazcami.
- `Types.h` – Obsahuje deklarácie používaných dátových typov a abstraktných tried reprezentujúcich rozhrania služieb zásuvných modulov.

Ďalej obsahuje 2 podadresáre:

- `TinyXML` – Zdrojový kód XML parseru Tiny XML.
- `utf8` – Obsahuje pomocné funkcie na prevod UTF-8 na UTF-16 a naopak. Zdrojom je [35].
- `wce` - Obsahuje súbory potrebné pre portovanie HTTP serveru na platformu Windows Mobile.

11.3.3 Adresár `WikiData`

Adresár `WikiData` obsahuje zdrojový kód modulu `WikiData.dll`. Obsahuje nasledujúce súbory:

- `WikiData.cpp` - Obsahuje exportované symboly potrebné pre funkčnosť zásuvného modulu.
- `WikiContent.h` – Obsahuje triedu `WikiContent`.
- `WikiHelper.h` – Obsahuje pomocné triedy a funkcie.
- `WikiFile.h` – Obsahuje triedu `WikiFile`.
- `WikiTitleIndex.h` – Obsahuje triedu `WikiTitleIndex`.

Tento adresár obsahuje taktiež adresár `WzipDec` so zdrojovým kódom dekompresora archívnych súborov (s formátom popísaným v kapitole 7.1.2). Tento adresár obsahuje súbory:

- `WzipDec.h` – Obsahuje triedu `WZip`.
- Ostatné súbory tvoria implementáciu LZMA dekodéru. Zdrojom týchto súborov je LZMA SDK [24].

11.3.4 Adresár `FileSystem`

V adresári `FileSystem` sa nachádza zdrojový kód modulu `FileSystem.dll`. Obsahuje nasledujúce súbory:

- `FileSystem.cpp` – Obsahuje exportované symboly potrebné pre funkčnosť zásuvného modulu.
- `FileSystem.h` – Obsahuje triedu `FileSystem`.

11.4 Testy

Funkčnosť programu bola overená na nasledujúcich zariadeniach:

- emulátor Windows Mobile 6.0 Professional
- emulátor Windows Mobile 6.0 Classic
- emulátor Windows Mobile 5.0
- Hewlett-Packard iPAQ 214, Windows Mobile 6.0 Classic
- Fujitsu-Siemens Pocket LOOX N560, Windows Mobile 5.0

Program nefunguje na operačnom systéme Windows Mobile 2003 a nižšom. Program je možné skompilovať pre tento systém, avšak z neznámej príčiny ho nie je možné spustiť v emulátore systému Windows Mobile 2003.

12 Budúcnosť programu

V budúcnosti sa naskytujú veľké možnosti rozšírenia riešenia popisovaného v tejto práci. V tejto kapitole sú popísané najdôležitejšie z nich.

12.1 Šablóny

Systém MediaWiki podporuje vkladanie obsahu iných článkov do aktuálneho článku. Navyše podporuje aj nahradzovanie parametrov v týchto vložených článkoch. Najčastejšie sú vkladané špeciálne články, ktoré sa nazývajú šablóny. Avšak je možné vkladať obsah ľubovoľných článkov. Funkcie šablón doplňujú parsovacie funkcie, ktoré fungujú podobne ako šablóny s tým rozdielom, že pre ne neexistujú samostatné články.

Implementácia tejto funkcie prinesie do článkov zobrazovaných v programe najmä viaceré informačné a orientačné tabuľky.

12.2 Vyhľadávanie

Pre spríjemnenie a zefektívnenie práce s programom bude vhodné pridať ďalšie typy indexov. Napríklad veľmi vhodným by bol index umožňujúci vyhľadávanie podľa zhody vo vnútri názvu článku bez rozlišovania diakritiky a veľkosti písmen.

Pôvodný index názvov článkov bude používaný pre dotazy pri ktorých potrebujeme poznať presný názov článku[1]. Nové indexy budú môcť byť zase používané pri vyhľadávaní. Nevýhodou bude nárast dátových súborov o dáta nových indexov. Používanie tejto funkcie bude voliteľné.

Pre implementáciu nových indexov bude potrebné pridať do konverznej časti mechanizmus, ktorý vytvorí žiadané indexy a implementovať nové služby umožňujúce prístup k týmto indexom. Zmena samotného programu WikiReader bude len minimálna – bude potrebné pridať službe `WikiFile` podporu používania viacerých indexových služieb. V konfiguračnom súbore dátového súboru bude nastaviteľné, ktorá indexová služba má byť použitá v konkrétnej situácii.

[1] Príkladom takýchto situácií sú odkazy v článkoch.

12.3 Kategórie

Dôležitou súčasťou Wikipédie sú kategórie článkov. Článok je zaradený do konkrétnej kategórie, ak sa v ňom nachádza odkaz na túto kategóriu. Tj. ak článok obsahuje odkaz „`[[Category:Demo]]`", tak sa nachádza v kategórii „Demo". Každý článok sa môže nachádzať vo viacerých kategóriách.

Pre implementáciu tejto funkcie bude potrebná zmena programu WikiConvert. Informácie o kategóriách bude potrebné uložiť do dátového súboru.

12.4 Referencie

Systém MediaWiki podporuje referencie a citácie prostredníctvom tagu `<ref>`, ktorý označuje samotnú citáciu a tagu `<references/>`, ktorý označuje miesto, na ktoré sa majú vložiť zdroje, ktoré sa nachádzajú v článku.

Implementácia tejto funkcie umožní dohľadanie zdrojov, v ktorých sa dajú overiť a rozšíriť informácie obsiahnuté v článkoch.

12.5 Matematické vzorce

Systém MediaWiki podporuje matematické vzorce zapísané v jazyku TeX. Tieto vzorce sú uzavreté do tagu `<math>`.

Naskytujú sa viaceré možnosti implementácie zobrazovania matematických vzorcov, avšak najprijateľnejšou sa zdá byť vygenerovanie obrázkov zo vzorcov pri konverzii a ich uloženie v dátovom súbore.

12.6 Podpora ďalších informačných zdrojov

V budúcnosti sa plánuje pridať podporu pre zobrazovanie offline obsahu ľubovoľných webových stránok. Tieto stránky bude potrebné stiahnuť pomocou sťahovača webových stránok a skomprimovať vhodnou kompresiou (napríklad formáty 7z alebo zip).

Veľmi vhodné informačné zdroje na mobilnom zariadení sú napríklad rôzne online kuchárky, tutoriály, stránky s dokumentáciou alebo s inými referenčnými informáciami.

12.7 Podpora iných mobilných platforiem

V budúcnosti sa plánuje portovať program WikiReader na iné mobilné platformy.

Program WikiReader bol tvorený s ohľadom na relatívne jednoduchú portáciu na iné platformy. Platformovo závislé časti boli zabalené do tried. Jedná sa hlavne o časti manažéra zásuvných modulov, webového serveru a LZMA dekompresie.

13 Záver

Riešenie popisované v tejto práci je v súčasnej podobe dobre použiteľné pre offline prístup k dátam z encyklopédie Wikipédia. Všetky komponenty riešenia sú voľne dostupné, a preto je možné toto riešenie ponúkať užívateľom zdarma. V dobe písania tejto práce sa jedná o jediné nám známe riešenie svojho typu na platforme Windows Mobile, ktoré je zdarma a ponúka dobrú podporu formátovania.

Program WikiReader je dostatočne rýchly a spoľahlivý. Navyše je šetrný k spotrebe pamäte zariadenia, na ktorom beží. Podpora formátovania je na dobrej úrovni. Najväčším nedostatkom formátovania je chýbajúca podpora šablón, čo ochudobňuje možnosti navigácie medzi článkami.

Program WikiReader bol od začiatku navrhovaný ako modulárny a to sa nám relatívne dobre podarilo. Nevýhodou modularity programu je však teoreticky nižšia rýchlosť. Program však podľa praktických testov beží dostatočne rýchlo a tento teoretický dopad modularity na jeho výkon nie je problémom.

Nevýhodou programu WikiReader je nemožnosť použitia iného webového prehliadača ako je Internet Explorer Mobile. Tento problém sme neboli schopní vyriešiť. Nie je to však problém nášho riešenia, ale limitácia operačného systému Windows Mobile.

Rýchlosť konverzie databázových exportov programom WikiConvert je veľmi dobrá. Použitie jazyka Java pre implementáciu programu WikiConvert umožnilo jeho nezávislosť na použitom operačnom systéme. Kompresná časť je však stále závislá na operačnom systéme, čo znemožňuje použitie konverznej časti na iných operačných systémoch ako tých, pre ktoré je možné skompilovať program WZip.

Zvolený spôsob kompresie algoritmom LZMA dosahuje veľmi dobré kompresné pomery a navyše umožňuje rýchlu dekompresiu. Nevýhodou vytvoreného kompresného formátu je však jeho nízka robustnosť. Napríklad v prípade poškodenia časti súboru s adresárom blokov súbor nie je ďalej použiteľný. Problémom je aj nemožnosť overenia integrity dát, pretože zvolený formát neobsahuje kontrolné súčty.

Zdroje

Nasledujúce zdroje som používal pri tvorbe tejto práce. Ich autorom preto ďakujem.

[1] ADAMS, Douglas. *Hitchhiker's guide to the galaxy*. [s.l.] : Pan Macmillan, 1979. 180 s. ISBN 0-330-25864-8.

[2] Yadabyte. *TomeRaider Ebook Reader* [online]. c2010 [cit. 2010-07-12]. Dostupné na: <http://tomeraider.com>.

[3] ZACHTE, Erik. *Complete Wikipedia on your handheld or notebook in TomeRaider format* [online]. 2003 [cit. 2010-07-12]. Dostupné na: <http://infodisiac.com/Wikipedia/>.

[4] Octopus Studio. *Octopus Studio* [online]. c2002 [cit. 2010-07-12]. Dostupné na: <http://www.octopus-studio.com>.

[5] ClearBits. *ClearBits* [online]. c2010 [cit. 2010-07-12]. Dostupné na: <http://www.clearbits.net>.

[6] *WikiPock* [online]. c2010 [cit. 2010-07-12]. Wikipedia for mobile phone. Dostupné na: <http://www.wikipock.com>.

[7] Linterweb. *Okawix* [online]. 2009 [cit. 2010-07-12]. Dostupné na: <http://www.okawix.com>.

[8] *Wiki2touch-standalone-ui* [online]. 2008 [cit. 2010-07-12]. Dostupné na: <http://code.google.com/p/wiki2touch-standalone-ui>.

[9] Wikimedia Foundation, Inc. *Wikimedia Foundation* [online]. c2003 [cit. 2010-07-12]. Home. Dostupné na: <http://wikimediafoundation.org>.

[10] Wikimedia Foundation, Inc. *Wikipedia, the free encyclopedia* [online]. August 2002, last modified on 5 July 2010 [cit. 2010-07-12]. Wikipedia. Dostupné na: <http://en.wikipedia.org/wiki/Wikipedia#Nature_of_Wikipedia>.

[11] Free Software Foundation, Inc. *GNU Project* [online]. Version 3. 29 June 2007 [cit. 2010-07-12]. The GNU General Public Licence. Dostupné na: <www.gnu.org/licenses/gpl.html>.

[12] Wikimedia Foundation, Inc. *MediaWiki* [online]. December 2005, last modified on 13 June 2010 [cit. 2010-07-12]. How does MediaWiki work?. Dostupné na: <http://www.mediawiki.org/wiki/How_does_MediaWiki_work?>.

[13] Creative Commons. *Creative Commons* [online]. 2009? [cit. 2010-07-12]. Attribution-ShareAlike 3.0 Unported. Dostupné na: <http://creativecommons.org/licenses/by-sa/3.0/legalcode>.

[14] Free Software Foundation, Inc. *GNU Project* [online]. Version 1.3. 3 November 2008 [cit. 2010-07-12]. GNU Free Documentation License. Dostupné na: <http://www.gnu.org/copyleft/fdl.html>.

[15] Wikimedia Foundation, Inc. *Wikimedia Downloads* [online]. 2010? [cit. 2010-07-12]. Database dump progress. Dostupné na: <http://dumps.wikimedia.org/backup-index.html>.

[16] Wikimedia Foundation, Inc. *MediaWiki* [online]. 2010? [cit. 2010-07-12]. Database export XML schema. Dostupné na: <http://www.mediawiki.org/xml/export-0.3.xsd>.

[17] Wikimedia Foundation, Inc. *MediaWiki* [online]. February 2006, last modified on 14 June 2010 [cit. 2010-07-12]. Download. Dostupné na: <http://www.mediawiki.org/wiki/Download>.

[18] Wikimedia Foundation, Inc. *MediaWiki* [online]. July 2010, last modified on 21 June 2010 [cit. 2010-07-12]. Manual:Importing XML dumps. Dostupné na: <http://www.mediawiki.org/wiki/Manual:Importing_XML_dumps>.

[19] PAVLOV, Igor. *7-Zip* [online]. c2009 [cit. 2010-07-12]. Dostupné na: <http://www.7-zip.org>.

[20] PKWARE Inc. *Pkware* [online]. Version 6.3.0. 2006-09-29 [cit. 2010-07-12]. .ZIP File Format Specification. Dostupné na: <http://www.pkware.com/documents/APPNOTE/APPNOTE-6.3.0.TXT>.

[21] *IEFT Tools* [online]. May 1996 [cit. 2010-07-12]. RFC 1952. Dostupné na: <http://tools.ietf.org/html/rfc1952>.

[22] SEWARD, Julian. bzip2 [online]. c1996 [cit. 2010-07-12]. Dostupné na: <http://bzip.org>.

[23] LÁNSKÝ, Jan. *Syllable-based Compression*. Praha, 2008. 123 s. Dizertačná práca. Univerzita Karlova, Matematicko-fyzikální fakulta. Dostupné online na: <http://www.ksi.mff.cuni.cz/~lansky/SC/disertacka.pdf>.

[24] PAVLOV, Igor. *7-Zip* [online]. c2009 [cit. 2010-07-12]. LZMA SDK. Dostupné na: <http://www.7-zip.org/sdk.html>.

[25] POKORNÝ, Jaroslav; ŽEMLIČKA, Michal. *Základy implementace souborů a databází*. 2. vydání. Praha : Nakladatelství Karolinum, 2004. 211 s. ISBN 80-246-0837-5. s. 21-25.

[26] Sun Microsystems, Inc. *Sun Microsystems* [online]. c1995 [cit. 2010-07-12]. Java Downloads for All Operating Systems. Dostupné na: <http://www.java.com/en/download/manual.jsp>.

[27] LARBI, Sam. *CodeOdor* [online]. 14 May 2007 [cit. 2010-07-12]. Re: Sorting really BIG files - the Java source code. Dostupné na: <http://www.codeodor.com/index.cfm/2007/5/14/Re-Sorting-really-BIG-files---the-Java-source-code/1208>.

[28] THOMASON, Lee. *TinyXml* [online]. 2001? [cit. 2010-07-12]. Dostupné na: <http://www.grinninglizard.com/tinyxml/>.

[29] NYFFENEGGER, René. *René Nyffenegger's collection of things on the web* [online]. c2003 [cit. 2010-07-12]. A Simple Webserver in C++. Dostupné na: <http://www.adp-gmbh.ch/win/misc/webserver.html>.

[30] *SourceForge.net* [online]. 2006 [cit. 2010-07-12]. Windows CE C Library Extensions. Dostupné na: <http://sourceforge.net/projects/wcelibcex/>.

[31] *IETF Tools* [online]. June 1999 [cit. 2010-07-12]. RFC 2616. Dostupné na: <http://tools.ietf.org/html/rfc2616>.

[32] Wikimedia Foundation, Inc. *Meta, a Wikimedia project coordination wiki* [online]. December 2003, last modified on 25 January 2010 [cit. 2010-07-12]. Help:Contents. Dostupné na: <http://meta.wikimedia.org/wiki/Help:Contents>.

[33] Microsof Corporation. *Microsoft Corporation* [online]. Version 1.0. 2007-05-01 [cit. 2010-07-12]. Windows Mobile 6 SDK Refresh. Dostupné z WWW: <http://www.microsoft.com/downloads/details.aspx?FamilyID=06111a3a-a651-4745-88ef-3d48091a390b&DisplayLang=en>.

[34] Microsof Corporation. *Microsoft Corporation* [online]. Version 5.0.14343. 2009-08-19 [cit. 2010-07-12]. Windows Mobile 5.0 SDK for Pocket PC. Dostupné z WWW: <http://www.microsoft.com/downloads/details.aspx?FamilyID=83a52af2-f524-4ec5-9155-717cbe5d25ed&DisplayLang=en>.

[35] *SourceForge.net* [online]. c2006 [cit. 2010-07-12]. UTF-8 CPP. Dostupné na: <http://sourceforge.net/projects/utfcpp/>.

Zoznam tabuliek

Zoznam diagramov

Príloha A – Popis priloženého DVD

/

- `bac.pdf` – Elektronická verzia bakalárskej práce.
- `wmemul.html` – Postup inštalácie samostatného emulátoru systému Windows Mobile.
- `demo.avi` – Video demonštrujúce základné funkcie programu WikiReader.

/bin/

Obsahuje spustiteľné súbory. Tento adresár obsahuje nasledujúce podadresáre:

- `WikiConvert` – Obsahuje spustiteľné súbory programu `WikiConvert`.
- `WikiReader` – Obsahuje spustiteľné súbory programu `WikiReader`.

/data/

Obsahuje ukážkové dátové súbory a súbor `config.xml`, ktorý predstavuje ukážkový konfiguračný súbor. Ďalej obsahuje tento adresár nasledujúce podadresáre:

- `cswiki` – Obsahuje dátové súbory českej Wikipédie.
- `simplewiki` – Obsahuje dátové súbory zjednodušenej anglickej Wikipédie.
- `skwiki` – Obsahuje dátové súbory slovenskej Wikipédie.

Z dôvodu nedostatočnej kapacity DVD média nie sú priložené dátové súbory anglickej Wikipédie.

/src/

Obsahuje zdrojový kód jednotlivých aplikácií popisovaných v práci. Obsahuje nasledujúce podadresáre:

- `WikiConvert` – Obsahuje zdrojový kód programu `WikiConvert`.
- `WikiReader` – Obsahuje zdrojový kód programu `WikiReader`.
- `WZip` – Obsahuje zdrojový kód programu `WZip`.
- `WikiStatistics` – Obsahuje zdrojový kód programu `WikiStatistics`.

Príloha B – Ukážka článku z databázového exportu

```
<page>
    <title>Alan Turing</title>
    <id>13</id>
    <revision>
      <id>1733111</id>
      <timestamp>2009-09-11T13:59:22Z</timestamp>
      <contributor>
        <username>Griffinofwales</username>
        <id>62069</id>
      </contributor>
      <minor />
      <comment>unsourced</comment>
      <text xml:space="preserve">'''Alan Mathison Turing''' ([[June 23]],
[[1912]] - [[June 7]], [[1954]]) was an [[England|English]] [[mathematician]]
and [[computer scientist]].

He was one of the first people to work with modern digital [[computer]]s. He
was the first person to think of using a computer for different things. He
told people that computers could run different [[Computer program|programs]].
Turing introduced the idea of a [[Turing machine]] in 1936. The machine was
imaginary, and ran a set of commands.

Turing also thought of the [[Turing test]].

During the [[Second World War]], Turing was a main participant in the efforts
to break German [[cipher]]s. On the basis of [[cryptanalysis]] he helped to
break both the [[Enigma machine]] and the [[Lorenz SZ 40/42]] (a teletype cip-
her attachment codenamed &amp;quot;Tunny&amp;quot; by the British), and was, for a ti-
me, head of [[Hut 8]], the section responsible for reading [[Germany|German]]
naval signals.

Alan Turing was a [[gay]] man. In [[1952]], Turing admitted having sex with a
man. At that time in [[England]], [[homosexuality]] was a crime. He was tried
and convicted of this crime in a British court. and was forced to make a choi-
ce. He had to choose between going to jail or &amp;quot;chemical castration&amp;quot;
(taking female [[hormone]]s like [[estrogen]] to lower his sex drive).  He
chose the hormones. But this made him impotent (unable to have sex) and made
him grow [[breast]]s. After suffering these effects for two years, he commit-
ted [[suicide]] (killed himself) with an [[apple]] poisoned with [[cyanide]]
in [[1954]].

The treatment forced on him is now believed to be very wrong, going against
[[medical ethics]] and international laws of [[human rights]], and [[malprac-
tice]] by most [[doctor]]s.

[[Category:English mathematicians|Turing, Alan]]
[[Category:Computer scientists|Turing, Alan]]
[[Category:LGBT people|Turing, Alan]]
[[Category:1912 births|Turing, Alan]]
[[Category:1954 deaths|Turing, Alan]]

{{Link FA|es}}
</text>
</revision>
</page>
```

Poznámka: Pri konverzii sa ukladá iba obsah tagu `<title>` a obsah tagu `<text>`.

Príloha C – Popis konfiguračných súborov

Konfiguračné súbory sú vo formáte XML. Popis týchto konfiguračných súborov obsahuje:

- **Identifikátor položky konfiguračného súboru**. Tento identifikátor obsahuje názvy jednotlivých „rodičovských“ položiek oddelených znakom „/“ a názov tejto položky. Neobsahuje názov koreňového elementu XML súboru. Identifikátory v popise obsahujú taktiež časti uzavreté do „< >“. Tieto časti reprezentujú variabilnú položku identifikátoru. V konfiguračnom súbore je potrebné tieto položky doplniť vhodným reťazcom.
- **Typ obsahu**, označuje dátový typ, ktorý sa môže vyskytovať v danej položke. Sú používané nasledujúce dátové typy:
  - prirodzené číslo
  - reťazec
  - URL adresa - Môže sa jednať aj o adresu relatívnu k adrese HTTP serveru programu WikiReader.
  - relatívna URL – Je relatívna k URL HTTP serveru programu WikiReader. Musí začínať znakom „/“.
  - cesta k súboru – Môžu byť použité 2 typu ciest k súborom:
    - úplná, ktorá začína znakom „\“
    - relatívna k aktuálnemu adresáru, v ktorom sa nachádza konfiguračný súbor. Môže začínať znakmi „**.** \”.
- **Popis položky**, obsahuje význam danej položky konfiguračného súboru.
- **Implicitná hodnota**, obsahuje implicitnú hodnotu, ktorá je použitá v programe ak sa daná položka nenachádza v konfiguračnom súbore. Ak obsahuje hodnotu „−“, tak je daná položka povinná.

Konfiguračný súbor programu WikiReader

V nasledujúcej tabuľke sa nachádza popis položiek v konfiguračnom súbore programu WikiReader.

Identifikátor položky	Typ obsahu	Popis položky	Imp.
`/server/threads`	prirodzené číslo	Maximálny počet vlákien HTTP serveru.	0
`/server/port`	prirodzené číslo z intervalu <0;65535>	Port na ktorom beží HTTP server.	80
`/style`	URL	URL CSS štýlu	“”

/plugin/<NS[1]>/src	cesta k súboru	Cesta k zásuvnému modulu, služby s názvom NS.	-
/plugin/<NS[1]>/service	reťazec	Identifikátor služby s názvom NS v rámci zásuvného modulu.	-
/data/<NDS[2]>/src	cesta k súboru	Cesta ku konfiguračnému súboru dátového súboru s názvom NDS.	-
/data/<NDS[1]>/mountdir	relatívna URL	Relatívna URL, ktorá označuje adresár serveru, z ktorého bude dátový súbor s názvom NDS prístupný	-
/data/<NDS[1]>/plugin	reťazec	Názov služby typu súborový formát, ktorá slúži pre prístup k dátovému súboru s názvom NDS.	-

Ukážkový konfiguračný súbor:

```
<?xml version="1.0" encoding="utf-8"?>
<config>
  <server>
    <threads>2</threads>
    <port>80</port>
  </server>
  <style>/sys/skin.css</style>
  <plugin>
    <WikiFile2>
      <src>WikiData.dll</src>
      <service>WikiFile</service>
    </WikiFile2>
  </plugin>
  <data>
    <sk>
      <src>sk\sk.xml</src>
      <mountdir>/sk/</mountdir>
      <plugin>WikiFile2</plugin>
    </sk>
    <en>
      <src>.\en\en.xml</src>
      <mountdir>/</mountdir>
      <plugin>WikiFile2</plugin>
    </en>
  </data>
</config>
```

Konfiguračný súbor dátových súborov Wikipédie

V nasledujúcej tabuľke sa nachádza popis položiek v konfiguračnom súbore dátových súborov, ktoré sú popisované v kapitole 7.1.

Identifikátor položky	Typ obsahu	Popis položky	Imp.
/content	reťazec	Názov služby typu konverzia obsahu, ktorá slúži na konverziu wiki článkov.	-
/data/volumes	prirodzené číslo z intervalu <1;255>	Počet zväzkov dátového archívu.	-
/data/src	cesta k súboru	Cesta k súboru dátového archívu.	-
/data/plugin	reťazec	Názov služby typu dekompresia, ktorá slúži na dekompresiu dátového archívu.	-
/index/title/src	cesta k súboru	Cesta k súboru s indexom názvov článkov.	-
/index/title/plugin	reťazec	Názov služby typu index, ktorá sluzi na vyhľadávanie v indexe názvov článkov.	-
/sitename	reťazec	Názov wiki projektu, ktorý sa nachádza v dátovom súbore.	""
/basename	reťazec	Odkaz na hlavnú stránku wiki projektu,	""

[1] Názov služby.
[2] Názov dátového súboru. Používaný ako identifikátor dátového súboru v programe.

		ktorý sa nachádza v dátovom súbore.	
/case	reťazec	Nastavenie citlivosti na veľkosť písmen názvov článkov. Ak obsahuje reťazec „first-letter“, tak je názov citlivý na veľkosť písmen až na prvé písmeno. Inak je citlivý na veľkosť písmen v celom názve.	“”
/namespaces/<nsid>[1]	reťazec	Lokalizovaný názov menného priestoru s identifikátorom nsid.	“”
/contains	reťazec	Obsahuje identifikátory menných priestorov, ktoré sa nachádzajú v dátovom súbore. Tieto identifikátory sú oddelené medzerami.	“”
/articles	prirodzené číslo	Počet článkov vyskytujúcich sa v dátovom súbore.	0

Ukážkový konfiguračný súbor:

```
<?xml version="1.0" encoding="utf-8"?>
<WikiFile>
  <content>WikiContent</content>
  <data>
    <volumes>5</volumes>
    <src>enwiki-20100130-pages-articles.w</src>
    <plugin>WZip</plugin>
  </data>
  <index>
    <title>
      <src>enwiki-20100130-pages-articles.index</src>
      <plugin>WikiTitleIndex</plugin>
    </title>
  </index>
  <sitename>Wikipedia</sitename>
  <basename>http://en.wikipedia.org/wiki/Main_Page</basename>
  <case>first-letter</case>
  <namespaces>
    <ns10>Template</ns10>
    <ns0></ns0>
    <ns14>Category</ns14>
  </namespaces>
  <contains>ns0 ns10</contains>
  <articles>7568171</articles>
</WikiFile>
```

[1] Identifikátor menného priestoru. V tvare ns9, kde číslo „9“ reprezentuje číselný identifikátor menného priestoru.